Speech Enhancement
Exploiting the Source-Filter Model

Von der Fakultät für Elektrotechnik, Informationstechnik, Physik
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte

Dissertation
(Kumulative Arbeit)

von
Samy Elshamy
aus Braunschweig

eingereicht am:	19.12.2019
mündliche Prüfung am:	29.05.2020
1. Referent:	Prof. Dr.-Ing. Tim Fingscheidt TU Carolo-Wilhelmina zu Braunschweig
2. Referent:	Prof. Dr.-Ing. Rainer Martin Ruhr-Universität Bochum
Vorsitz:	Prof. Dr.-Ing. Eduard A. Jorswieck TU Carolo-Wilhelmina zu Braunschweig
Druckjahr:	2020

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

Abstract

Imagining everyday life without mobile telephony is nowadays hardly possible. Calls are being made in every thinkable situation and environment. Hence, the microphone will not only pick up the user's speech but also sound from the surroundings which is likely to impede the understanding of the conversational partner. Modern speech enhancement systems are able to mitigate such effects and most users are not even aware of their existence.

In this thesis the development of a modern single-channel speech enhancement approach is presented, which uses the divide and conquer principle to combat environmental noise in microphone signals. Though initially motivated by mobile telephony applications, this approach can be applied whenever speech is to be retrieved from a corrupted signal. The approach uses the so-called source-filter model to divide the problem into two subproblems which are then subsequently conquered by enhancing the source (the excitation signal) and the filter (the spectral envelope) separately. Both enhanced signals are then used to denoise the corrupted signal. The estimation of spectral envelopes has quite some history and some approaches already exist for speech enhancement. However, they typically neglect the excitation signal which leads to the inability of enhancing the fine structure properly. Both individual enhancement approaches exploit benefits of the cepstral domain which offers, e.g., advantageous mathematical properties and straightforward synthesis of excitation-like signals.

We investigate traditional model-based schemes like Gaussian mixture models (GMMs), classical signal processing-based, as well as modern deep neural network (DNN)-based approaches in this thesis. The enhanced signals are not used directly to enhance the corrupted signal (e.g., to synthesize a clean speech signal) but as so-called *a priori* signal-to-noise ratio (SNR) estimate in a traditional statistical speech enhancement system. Such a traditional system consists of a noise power estimator, an *a priori* SNR estimator, and a spectral weighting rule that is usually driven by the results of the aforementioned estimators and subsequently employed to retrieve the clean speech estimate from the noisy observation.

As a result the new approach obtains significantly higher noise attenuation compared to current state-of-the-art systems while maintaining a quite comparable speech component quality and speech intelligibility. In consequence, the overall quality of the enhanced speech

https://doi.org/10.24355/dbbs.084-202102121510-0

signal turns out to be superior as compared to state-of-the-art speech ehnahcement approaches.

https://doi.org/10.24355/dbbs.084-202102121510-0

Zusammenfassung

Mobiltelefonie ist aus dem heutigen Leben nicht mehr wegzudenken. Telefonate werden in beliebigen Situationen an beliebigen Orten geführt und dabei nimmt das Mikrofon nicht nur die Sprache des Nutzers auf, sondern auch die Umgebungsgeräusche, welche das Verständnis des Gesprächspartners stark beeinflussen können. Moderne Systeme können durch Sprachverbesserungsalgorithmen solchen Effekten entgegenwirken, dabei ist vielen Nutzern nicht einmal bewusst, dass diese Algorithmen existieren.

In dieser Arbeit wird die Entwicklung eines einkanaligen Sprachverbesserungssystems vorgestellt. Der Ansatz setzt auf das Teile-und-herrsche-Verfahren, um störende Umgebungsgeräusche aus Mikrofonsignalen herauszufiltern. Dieses Verfahren kann für sämtliche Fälle angewendet werden, in denen Sprache aus verrauschten Signalen extrahiert werden soll. Der Ansatz nutzt das Quelle-Filter-Modell, um das ursprüngliche Problem in zwei Unterprobleme aufzuteilen, die anschließend gelöst werden, indem die Quelle (das Anregungssignal) und das Filter (die spektrale Einhüllende) separat verbessert werden. Die verbesserten Signale werden gemeinsam genutzt, um das gestörte Mikrofonsignal zu entrauschen. Die Schätzung von spektralen Einhüllenden wurde bereits in der Vergangenheit erforscht und zum Teil auch für die Sprachverbesserung angewandt. Typischerweise wird dabei jedoch das Anregungssignal vernachlässigt, so dass die spektrale Feinstruktur des Mikrofonsignals nicht verbessert werden kann. Beide Ansätze nutzen jeweils die Eigenschaften der cepstralen Domäne, die unter anderem vorteilhafte mathematische Eigenschaften mit sich bringen, sowie die Möglichkeit, Prototypen eines Anregungssignals zu erzeugen.

Wir untersuchen modellbasierte Ansätze, wie z.B. Gaußsche Mischmodelle, klassische signalverarbeitungsbasierte Lösungen und auch moderne tiefe neuronale Netzwerke in dieser Arbeit. Die so verbesserten Signale werden nicht direkt zur Sprachsignalverbesserung genutzt (z.B. Sprachsynthese), sondern als sogenannter A-priori-Signal-zu-Rauschleistungs-Schätzwert in einem traditionellen statistischen Sprachverbesserungssystem. Dieses besteht aus einem Störleistungs-Schätzer, einem A-priori-Signal-zu-Rauschleistungs-Schätzer und einer spektralen Gewichtungsregel, die üblicherweise mit Hilfe der Ergebnisse der beiden Schätzer berechnet wird. Schließlich wird eine Schätzung des sauberen Sprachsignals aus der Mikrofonaufnahme gewonnen.

Der neue Ansatz bietet eine signifikant höhere Dämpfung des Störgeräuschs als der bisherige

https://doi.org/10.24355/dbbs.084-202102121510-0

Stand der Technik. Dabei wird eine vergleichbare Qualität der Sprachkomponente und der Sprachverständlichkeit gewährleistet. Somit konnte die Gesamtqualität des verbesserten Sprachsignals gegenüber dem Stand der Technik erhöht werden.

https://doi.org/10.24355/dbbs.084-202102121510-0

Contents

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

List of Publications

I S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "An Iterative Speech Model-Based A Priori SNR Estimator," in *Proc. of Interspeech*, Dresden, Germany, Sep. 2015, pp. 1740–1744

II S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "Instantaneous A Priori SNR Estimation by Cepstral Excitation Manipulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017

III S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "Two-Stage Speech Enhancement with Manipulation of the Cepstral Excitation," in *Proc. of HSCMA*, San Francisco, CA, USA, Mar. 2017, pp. 106–110

IV S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "DNN-Supported Speech Enhancement With Cepstral Estimation of Both Excitation and Envelope," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2460–2474, Dec. 2018

V S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "A Priori SNR Computation for Speech Enhancement Based on Cepstral Envelope Estimation," in *Proc. of IWAENC*, Tokyo, Japan, Sep. 2018, pp. 531–535

VI S. Elshamy and T. Fingscheidt, "DNN-Based Cepstral Excitation Manipulation for Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1803–1814, Nov. 2019

VII S. Elshamy and T. Fingscheidt, "Improvement of Speech Residuals for Speech Enhancement," in *Proc. of WASPAA*, New Paltz, NY, USA, Oct. 2019, pp. 214–218

https://doi.org/10.24355/dbbs.084-202102121510-0

Please note that the order of the publications listed above is corresponding to the chronological order of their writing, which is not necessarily coherent with the publication dates. This is important as the publications are based on each other. However, Publications **II**, **III**, **VI**, and **VII** form a thematic complex, Publications **IV** and **V** another, and Publication **I** is to be seen as an independent preliminary study. The structure of this thesis and the order of publications therein follows exactly this thematic clustering.

https://doi.org/10.24355/dbbs.084-202102121510-0

Author's Contribution

I The algorithm was implemented by the author and all corresponding experiments and analyses were conducted by the author. The adaptation of the existing framework [Mowlaee and Saeidi, 2013] for *a priori* SNR estimation was an idea of Nilesh Madhu. The setup of the experiments was jointly designed by the author and the co-authors. The author was primarily responsible for the writing of the paper.

II The algorithm was implemented by the author and all corresponding experiments and analyses were conducted by the author. The idea to use the source-filter model and enhance either component separately was a result of various analyses and discussions of the author and the co-authors. The core manipulations of the proposed approach were an idea of the author. The setup of the experiments was jointly designed by the author and the co-authors. The author was primarily responsible for the writing of the article. The writing of Sections V-C and D was partly done by Tim Fingscheidt.

III The algorithm was implemented by the author and all corresponding experiments and analyses were conducted by the author. The idea to use the source-filter model and enhance either component separately was a result of various analyses and discussions of the author and the co-authors. The core manipulations of the proposed approach were an idea of the author. The setup of the experiments was jointly designed by the author and the co-authors. The author was primarily responsible for the writing of the paper.

https://doi.org/10.24355/dbbs.084-202102121510-0

IV The algorithm was implemented by the author and all corresponding experiments and analyses were conducted by the author. The idea to use a hidden Markov model (HMM) with all its possibilities was a joint effort of the co-authors. The serial concatenation of the proposed approaches was an idea of Wouter Tirry. Setup, experiments, and analyses were designed and conducted by the author who was also primarily responsible for the writing of the article.

V The algorithm was implemented by the author and all corresponding experiments and analyses were conducted by the author. The author was primarily responsible for the writing of the paper.

VI The algorithm was implemented by the author and all corresponding experiments and analyses were conducted by the author. To investigate a second target representation was an idea of Tim Fingscheidt. The author was primarily responsible for the writing of the article.

VII The algorithm was implemented by the author and all corresponding experiments and analyses were conducted by the author. The subjective listening test setup was designed, conducted, and evaluated by the author. The author was primarily responsible for the writing of the paper.

https://doi.org/10.24355/dbbs.084-202102121510-0

List of Abbreviations

ABE	Artificial bandwidth extension
ASR	Automatic speech recognition
CCR	Comparison category rating
CEE	Cepstral envelope enhancement
CEM	Cepstral excitation manipulation
CMOS	Comparison mean opinion score
CTS	Cepstro-temporal smoothing
DD	Decision-directed
DFT	Discrete Fourier transform
DNN	Deep neural network
FoM	Figure of merit
GMM	Gaussian mixture model
HMM	Hidden Markov model
HRNR	Harmonic regeneration noise reduction
IDFT	Inverse discrete Fourier transform
LPC	Linear predictive coding
LQO	Listening quality objective
MAP	Maximum a posteriori
MB	Model-based
ML	Maximum likelihood
MMSE	Minimum mean squared error
MMSE-LSA	Minimum mean squared error log-spectral amplitude
MMSE-STSA	Minimum mean squared error short-time spectral amplitude
MOS	Mean opinion score
NA	Noise attenuation
NPE	Noise power estimator
NTT	Nippon Telegraph and Telephone
PESQ	Perceptual evaluation of speech quality
RMS	Root-mean-square
SG-jMAP	Super-Gaussian joint maximum *a posteriori*

https://doi.org/10.24355/dbbs.084-202102121510-0

SNR	Signal-to-noise ratio
SSDR	Speech-to-speech distortion ratio
STFT	Short-time Fourier transform
STOI	Short-time objective intelligibility
WF	Wiener filter
WR	Weighting rule

https://doi.org/10.24355/dbbs.084-202102121510-0

List of Symbols

$A_\ell(k)$	LPC coefficients spectrum
$D_\ell(k)$	Noise signal spectrum
$G_\ell(k)$	Spectral weighting rule
$H_\ell(k)$	Envelope spectrum
K	DFT size
$R_\ell(k)$	Residual signal spectrum
$S_\ell(k)$	Clean speech signal spectrum
$Y_\ell(k)$	Microphone signal spectrum
ℓ	Frame index
$\hat{S}_\ell(k)$	Enhanced speech signal spectrum
$\hat{\gamma}_\ell(k)$	*A posteriori* SNR
$\hat{\sigma}_\ell^D(k)^2$	Noise power estimate
$\hat{\xi}_\ell(k)$	*A priori* SNR
$\hat{s}(n)$	Enhanced speech signal
$a(i)$	LPC coefficients
$c_\ell^H(m)$	Envelope cepstrum
$c_\ell^R(m)$	Residual signal cepstrum
$d(n)$	Noise signal
f_0	Fundamental frequency
k	Frequency bin index
m	Quefrency bin index
n	Discrete-time sample index
$s(n)$	Clean speech signal
$y(n)$	Microphone signal

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

1 Introduction

Speech is one of the most important and intuitive means of communication for the majority of human beings. Modern telephony enables communication through various channels and renders conversation possible despite any distance or circumstance. However, a microphone is still required to capture the sound produced by a sending human being at the near end and to transport it accordingly to the receiver at the far end of the conversation. Thereby, new issues arise as the microphone picks up not only the desired speech but also any sound that arrives at the membrane. The presence of noise at the near end is likely to degrade the quality and also the intelligibility of the speech signal at the far end since speech and noise are jointly captured at the microphone. Thus, speech enhancement algorithms—more specifically noise reduction algorithms—are usually employed in the uplink as a countermeasure in order to denoise the microphone signal and retrieve the desired speech signal as clean as possible.

Further speech enhancement systems embrace, e.g., artificial bandwidth extension and acoustic echo cancellation, where the purpose is also to improve intelligibility, quality, and listening comfort for mobile communication devices, human-machine interfaces, and also the hearing impaired.

In this thesis we focus on noise reduction and depict the development of a novel model-based approach for communication systems. One novel aspect lies within the utilization of the source-filter model and the specific enhancement of both components. The source-filter model allows to decompose a given speech signal into a smooth spectral envelope (the filter) and the corresponding excitation signal (the source). Both components have specific properties due to the model constraint, which are beneficial for separate enhancement strategies that are investigated. A further aspect is the introduction of modern machine learning algorithms after traditional schemes are explored and evaluated subsequently. The developed method is embedded in a traditional common noise reduction framework consisting of a noise power estimator, an *a priori* signal-to-noise ratio (SNR) estimator, and a spectral weighting rule, where the proposed method is used for *a priori* SNR estimation.

Even though this thesis targets single-channel speech enhancement for narrowband telephony, an incorporation of, e.g., multi-channel noise power estimators or an adaptation towards telephony beyond narrowband should be straightforward. For noise reduction sys-

https://doi.org/10.24355/dbbs.084-202102121510-0

tems, there is usually a trade-off between noise attenuation and speech (component) quality or intelligibility. The novel approach manages to circumvent this trade-off to some extent and allows for a significantly higher noise attenuation while maintaining comparable speech component quality and intelligibility. Also the robustness against non-stationary noises and noises that are unknown to the system is shown. A subjective listening test supports the objective results and shows that the new method is preferred over current state-of-the-art systems with significant results.

Most of the research in this thesis has been conducted during a collaboration with NXP Software, where parts of this thesis have contributed to the `LifeVibes VoiceExperience` software suite. Accordingly, a patent for some of the proposed technology has been filed and granted in various countries [Elshamy et al., 2017b, Elshamy et al., 2019a, Elshamy et al., 2019b].

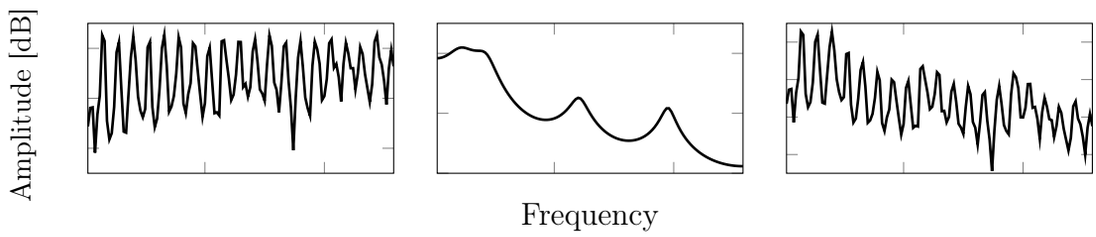
1.1 Source-Filter Model of Human Speech Production

Since one of the novel aspects of this thesis is the utilization of the source-filter model to split the problem into two subproblems, its origin and functioning are explained in this section.

The source-filter model of human speech production allows for a simplifying mathematical model of how speech is formed in human beings. It is understood as a two-stage interaction of lungs together with glottis as the first stage and vocal tract as the second stage. The lungs pump air through the glottis and depending on whether the vocal chords are vibrating or not, an either voiced or unvoiced excitation signal is generated, respectively [Flanagan, 1965]. These two types of excitation can be modeled either as a train of equidistantly spaced pulses, when the vocal chords are vibrating or as white noise otherwise. The excitation signal is then shaped by the vocal tract while flowing through it, before radiating through the lips. The parameters for such a model can be obtained, e.g., by linear predictive coding (LPC) analysis [Markel and Gray, 1976], where the vocal tract is often understood as an all-pole linear filter. Since this model is a simplification to also reduce complexity, some inaccuracies are inevitable which are, however, negligible [O'Shaughnessy, 1987].

Figure 1.1 visualizes the source-filter model by showing a voiced excitation signal (Figure 1.1a) as the source, the corresponding spectral envelope (Figure 1.1b) as the filter, and the resulting voiced speech signal (Figure 1.1c). The sample is obtained from the Nippon Telegraph and Telephone (NTT) super wideband database [NTT, 2012] downsampled to 8 kHz and the individual exhibits a fundamental frequency $f_0 \approx 200\,\text{Hz}$. One can see that the formerly "flat" source signal after being filtered by the vocal tract-representing spectral envelope obtains a shape that is similar to the spectral envelope. It is also reasonable to

https://doi.org/10.24355/dbbs.084-202102121510-0



(a) Voiced excitation signal (source)

(b) Spectral envelope (filter)

(c) Voiced speech signal

Figure 1.1: Example of the source-filter model by LPC analysis.

consider both components separately in speech enhancement as they both have their own distinct characteristics, which might facilitate the estimation of both signals separately, instead of estimating the joint speech signal.

Another way to obtain a spectral envelope and the corresponding excitation signal is the so-called liftering of a cepstrum. It is the analogue to filtering when considering a spectrum. In the cepstral domain, the lower-indexed coefficients are representing low-frequency waves of the analyzed spectrum. Now, these coefficients are generally attributed to the spectral envelope, while the higher-indexed coefficients are representing the fine structure of the spectrum and thus are attributed to the excitation signal. Then, to obtain the spectral envelope from the speech spectrum in the cepstral domain, the lower-indexed quefrency bins are isolated by cutting off the remaining quefrency bins. The excitation signal is obtained analogously. A special role plays the zeroth coefficient, which represents the energy level of the spectrum.

There are no constraints that ensure specific characteristics of both components. However, LPC analysis will always yield an excitation signal, which is shapeless in the sense that the spectral envelope will carry all the shaping information and the excitation signal appears to be flat in shape, while still being allowed to oscillate. Furthermore, the spectral envelope is limited to a fixed number of poles and is modeled as a filter when applying this method. An example is depicted in Figure 1.2, where it can be seen that the cepstral approach is delivering merely a smoothed version of the spectrum, which is exactly what the liftering (here of the first 31 coefficients) does, as it removes the high-frequency portions of the spectrum. This has been also analyzed and shown in [Benesty et al., 2008, Sec. 9.5.1]. Therefore, it is unsuitable for our method as it does not guarantee the components to have specific characteristics as when obtained by LPC analysis, where the envelope is being modeled as an all-pole filter and the excitation signal is rendered spectrally "flat". Having a certain kind of homogeneity is quite an important factor when models are to be learned, which is obtained by the mentioned characteristics in this case. Since filter coefficients are quite sensitive and

https://doi.org/10.24355/dbbs.084-202102121510-0

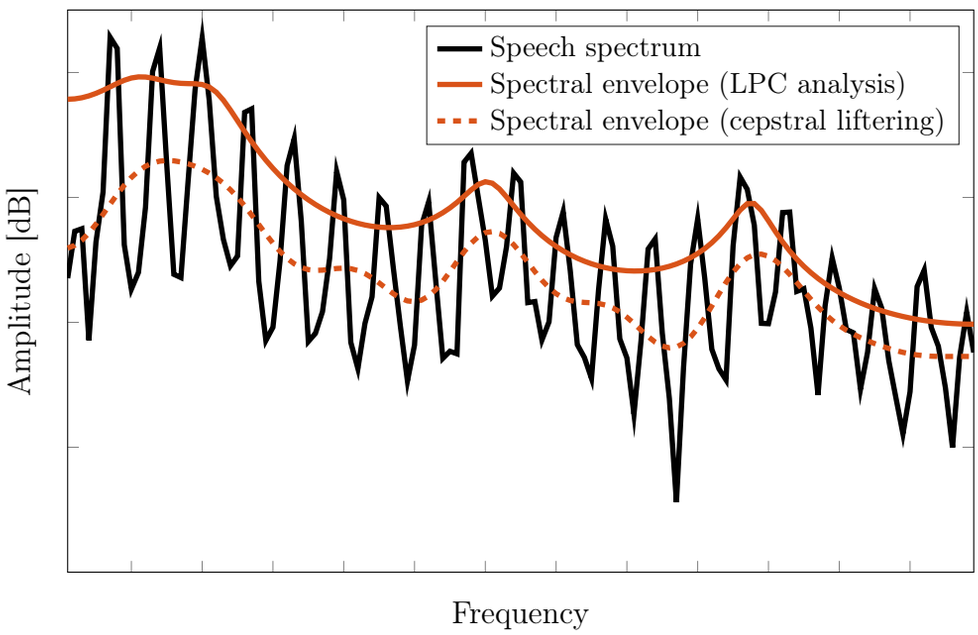


Figure 1.2: Example showing a speech spectrum and the corresponding spectral envelope obtained by 10th order LPC analysis and cepstral liftering with cut-off quefrency corresponding to approximately 500 Hz.

small manipulations might lead to instabilities, a transformation [Papamichalis, 1987] can be used to obtain a cepstral representation from LPC coefficients. This allows to obtain a representation of the spectral envelope that facilitates simple operations such as averaging, which might be necessary for certain training algorithms, while providing meaningful results.

1.2 Signal Model and Some Notations

In this section we introduce our signal model, important mathematical symbols, and some notations that are relevant for this thesis. We assume an additive model for the noisy time-domain microphone signal $y(n)$, so that for every discrete-time sample index n

$$y(n) = s(n) + d(n) \tag{1.1}$$

holds true. The clean speech time-domain signal is denoted as $s(n)$ and the time-domain noise signal as $d(n)$.

The corresponding frequency-domain representation is obtained by applying a K-Point

https://doi.org/10.24355/dbbs.084-202102121510-0

discrete Fourier transform (DFT) to the time-domain signals, yielding

$$Y_\ell(k) = S_\ell(k) + D_\ell(k). \tag{1.2}$$

Here, the frequency-domain microphone representation is $Y_\ell(k)$, the frequency-domain clean speech representation $S_\ell(k)$, and the frequency-domain noise representation $D_\ell(k)$. The frame index is denoted by ℓ and $0 \leq k \leq K-1$ represents the frequency bin index. Following common speech enhancement approaches [Ephraim and Malah, 1984, Wolfe and Godsill, 2001] we assume statistical independence of speech and noise as well as zero-mean signals. Even though this might not be strictly true in every case, it is a reasonable approach in practice [Stylianou et al., 2007].

Since we use LPC analysis to further evaluate some signals, the corresponding LPC coefficients are depicted by $a(i)$, $A_\ell(k)$ in the time and frequency domain, respectively. We refer to the spectral envelope as $H_\ell(k)$ and the corresponding residual signal spectrum as $R_\ell(k)$. A cepstral representation of the residual signal is denoted by $c_\ell^R(m)$ while its counterpart, the spectral envelope, is denoted by $c_\ell^H(m)$.

Traditional noise reduction algorithms require a per frame and bin-wise noise power estimate $\hat{\sigma}_\ell^D(k)^2$, an *a priori* SNR estimate $\hat{\xi}_\ell(k)$, optionally also the *a posteriori* SNR $\hat{\gamma}_\ell(k)$, to calculate a spectral weighting rule $G_\ell(k)$.

The aim of speech enhancement is to retrieve the enhanced clean speech signal as $\hat{s}(n)$ or $\hat{S}_\ell(k)$, in the time domain or frequency domain, respectively. This is traditionally done by applying the spectral weighting rule as

$$\hat{S}_\ell(k) = Y_\ell(k) \cdot G_\ell(k). \tag{1.3}$$

Estimated entities are generally denoted by the hat operator $\hat{\cdot}$ and superscripts are usually referring to a specific signal for the corresponding entity.

1.3 Evaluation Metrics

The evaluation of a speech enhancement system, especially a noise reduction algorithm, is crucial and also challenging. Its importance is indisputable as, e.g., speech quality, speech intelligibility, and also the attenuation of noise are essential aspects of such a system. We distinguish between objective and subjective measures, where the objective measures are algorithms that are able to deterministically rate enhanced signals under various aspects. Measures which require a reference signal to compute the score are called intrusive, while non-intrusive measures operate only on the enhanced signal. The use of such algorithms—when applied as intended—usually allows to develop new systems and obtain feedback

https://doi.org/10.24355/dbbs.084-202102121510-0

about its performance quickly. They are quite essential for research and development, even though being less accurate, since subjective measures always require careful preparing of time-consuming listening tests with human beings. The following will introduce all the objective measures and underlying principles that have been used in this thesis for evaluation.

Perceptual Evaluation of Speech Quality

Objective measures often aim to estimate or model subjective tests, e.g., the intrusive perceptual evaluation of speech quality (PESQ) measure [ITU, 2001, ITU, 2003] models the mean opinion score (MOS) [ITU, 1996] and also the MOS-listening quality objective (LQO), respectively. PESQ has been intentionally designed to rate the speech quality of narrowband speech codecs. However, it is widely used to rate enhanced speech signals, also processed by noise reduction algorithms, e.g., [Hendriks et al., 2010, Sigg et al., 2012, He et al., 2017, Huang et al., 2018]. This is, if at all, mostly based on the reported correlation between PESQ and subjective ratings of processed signals in [Hu and Loizou, 2008]. The recommendation specifying PESQ [ITU, 2001], however, states that PESQ has not been validated against artifacts or effects from noise reduction algorithms. For the sake of completeness we want to mention that the method has been extended to wideband in [ITU, 2007] and the raw MOS scores provided by [ITU, 2001] are mapped to MOS-LQO by [ITU, 2007], thereby mapping raw scores from $-0.5 \ldots 4.5$ to $1.02 \ldots 4.56$.

White-Box Approach

One way to mitigate the issue that PESQ has not been validated against artifacts or effects from noise reduction algorithms is to apply the so-called white-box approach [Gustafsson et al., 1996, Suhadi, 2012]. It allows to investigate the effects that a filter has on the separate components of a single mixed signal. The application of the white-box approach is only possible in a simulation environment where all components of the generated signals can be accessed separately and the components are superimposed to obtain the signal that is to be processed. A generic block diagram of the white-box approach for signals consisting of two components, here speech and noise, is shown in Figure 1.3.

For the presented example, following

$$Y_\ell(k) = S_\ell(k) + D_\ell(k) \tag{1.4}$$

and

$$\hat{S}_\ell(k) = Y_\ell(k) \cdot G_\ell(k), \tag{1.5}$$

https://doi.org/10.24355/dbbs.084-202102121510-0

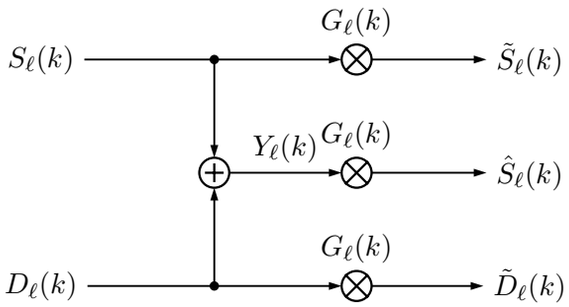


Figure 1.3: Diagram of the white-box approach showing how to obtain separately filtered signal components, here for two signals, after [Gustafsson et al., 1996].

it is evident that

$$\hat{S}_\ell(k) = (S_\ell(k) + D_\ell(k)) \cdot G_\ell(k). \tag{1.6}$$

This allows to obtain separately filtered components which in sum represent the microphone signal by

$$\hat{S}_\ell(k) = S_\ell(k) \cdot G_\ell(k) + D_\ell(k) \cdot G_\ell(k), \tag{1.7}$$

where the filtered components are obtained as

$$\tilde{S}_\ell(k) = S_\ell(k) \cdot G_\ell(k) \tag{1.8}$$

being the filtered speech component and

$$\tilde{D}_\ell(k) = D_\ell(k) \cdot G_\ell(k) \tag{1.9}$$

being the filtered noise component.

This method provides a way to assess the filtered components separately with corresponding measures. It allows to interpret the effects on the filtered speech component $\tilde{S}_\ell(k)$ similarly to coding distortions, which is then more in line with the intended use case of PESQ rather than measuring $\hat{S}_\ell(k)$.

Segmental Speech-to-Speech Distortion Ratio

A further measure to assess the speech component quality is the segmental speech-to-speech distortion ratio (SSDR) [Fingscheidt et al., 2008]. The segmental SSDR is not based on a perceptual model but is merely a sample by sample-comparing approach to rate the speech distortion of a processed signal w.r.t. to the corresponding clean reference signal, hence, being also an intrusive measure. A frame is considered to be a segment in the following. The metric is calculated as an average over all frames as

$$\mathrm{SSDR}_{\mathrm{seg}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \mathrm{SSDR}(\ell) \tag{1.10}$$

https://doi.org/10.24355/dbbs.084-202102121510-0

with frame index ℓ and $\mathcal{L}$ being the set that contains all frame indices of the signals that are to be considered. This allows to either evaluate all frames or, e.g., a subset including only speech active frames. The frame-based SSDR is usually limited to values between $R_{\min} = -10\,\text{dB}$ and $R_{\max} = 30\,\text{dB}$ by

$$\text{SSDR}(\ell) = \max\left\{\min\left\{\text{SSDR}'(\ell), R_{\max}\right\}, R_{\min}\right\}. \tag{1.11}$$

The logarithmic signal-to-error ratio is calculated for each frame as

$$\text{SSDR}'(\ell) = 10\log_{10}\left[\frac{\sum_{\nu=0}^{N-1} s(\nu+\ell N)^2}{\sum_{\nu=0}^{N-1} e(\nu+\ell N)^2}\right] \tag{1.12}$$

with N being the length of a frame and the sample-by-sample error is calculated as

$$e(\nu+\ell N) = \tilde{s}(\nu+\ell N) - s(\nu+\ell N). \tag{1.13}$$

It is important to mention that the SSDR requires both signals to be time-aligned after filtering. A high SSDR_{seg} is anticipated for a good speech enhancement system as it indicates a very low distortion of the processed signal compared to the reference.

Short-Time Objective Intelligibility

Besides speech quality, there is another factor that is quite important for speech enhancement systems. The intelligibility of speech is another important measure, which reflects the comprehensibility of speech independently of quality or other aspects. The short-time objective intelligibility (STOI) measure [Taal et al., 2011] is an objective, intrusive algorithm, that takes a processed signal and its corresponding clean reference and generates a scalar output in the range of $[0, 1]$. It has been explicitly tested for signals processed by conventional noise reduction systems and a value close to unity is to be anticipated.

Segmental Noise Attenuation

The second dimension of quality considers the amount of achieved noise suppression. It is important to consider both dimensions since it is easy to achieve a very high level of suppression, e.g., by using static broadband attenuation, thereby automatically affecting the speech component, which would go unnoticed if only one dimension is considered. Usually, for noise reduction algorithms there is a trade-off between speech quality or speech intelligibility on the one hand, and noise attenuation on the other hand, as the aggressiveness of a noise reduction algorithm easily tends to also attenuate and distort the speech component.

https://doi.org/10.24355/dbbs.084-202102121510-0

The segmental noise attenuation (NA) [Fingscheidt et al., 2008] is the average of a local frame-wise ratio of the noise power and the power of the filtered noise component. It is calculated as

$$\mathrm{NA}_{\mathrm{seg}} = 10 \log_{10} \left[\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \mathrm{NA}(\ell) \right], \tag{1.14}$$

where the set $\mathcal{L}$ allows to restrict the measure to certain, e.g., speech-active frames only. The ratio for each segment or frame is calculated as

$$\mathrm{NA}(\ell) = \frac{\sum_{\nu=0}^{N-1} d(\nu + \ell N)^2}{\sum_{\nu=0}^{N-1} \tilde{d}(\nu + \ell N)^2}, \tag{1.15}$$

with N being the length of a frame. The signals also have to be time-aligned after processing and a high $\mathrm{NA}_{\mathrm{seg}}$ reflects a high noise attenuation and is favored.

Delta Signal-to-Noise Ratio

Supplementary to the segmental NA, the delta SNR measure quantifies the SNR improvement on a global level. It is based on the active speech and root-mean-square (RMS) level measurements from ITU-T P.56 [ITU, 2011], where the first is used to measure the level of a speech signal and the second to measure the level of a noise signal. The delta SNR is then calculated as

$$\Delta\mathrm{SNR} = \mathrm{SNR}_{\mathrm{out}} - \mathrm{SNR}_{\mathrm{in}}. \tag{1.16}$$

Thereby, the output SNR after processing is depicted by $\mathrm{SNR}_{\mathrm{out}}$ and the input SNR prior to processing by $\mathrm{SNR}_{\mathrm{in}}$, correspondingly. This measure can also only be applied in a simulation environment since the clean speech and the superimposed noise are required for the SNR calculation of $\mathrm{SNR}_{\mathrm{in}}$. Furthermore, this method requires the filtered components to calculate $\mathrm{SNR}_{\mathrm{out}}$. Here, also a high ΔSNR value is of interest and represents good noise reduction performance.

1.4 State of the Art

Traditional speech enhancement algorithms for noise reduction consist of three main components. Such a traditional system is depicted in Figure 1.4 for a generic approach based in the frequency domain. Here, the first component is a noise power estimator (NPE), which calculates a noise power spectral density estimate $\hat{\sigma}_\ell^D(k)^2$ based on the microphone signal. A simple approach is to identify speech inactive frames and assume their power spectral density as the noise power estimate. More advanced approaches track spectral minima and use time-variant smoothing factors to obtain the desired estimate based on the microphone signal, e.g., [Martin, 2001, Cohen, 2003]. Furthermore, minimum mean squared error (MMSE)-based estimation has been proposed and shown that it can be interpreted

https://doi.org/10.24355/dbbs.084-202102121510-0

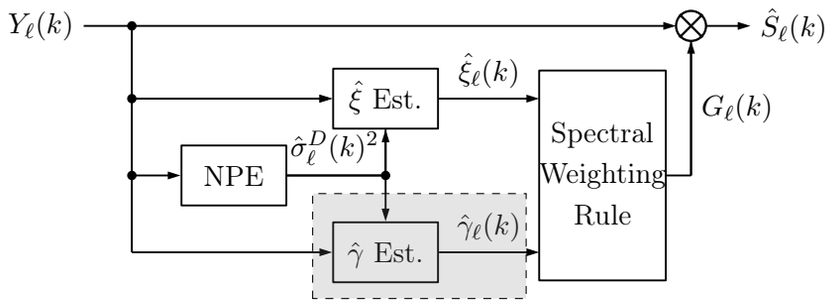


Figure 1.4: Diagram of a traditional noise reduction algorithm in the frequency domain with three main components: noise power estimator (NPE), *a priori* SNR ($\hat{\xi}_\ell(k)$) estimator (optionally also *a posteriori* SNR ($\hat{\gamma}_\ell(k)$) estimator), and spectral weighting rule.

as a voice activity detector [Gerkmann and Hendriks, 2012]. The noise power estimate $\hat{\sigma}_\ell^D(k)^2$ is then fed to the second component which is responsible for the estimation of the *a priori* SNR $\hat{\xi}_\ell(k)$. Optionally, in addition to that, the so-called *a posteriori* SNR $\hat{\gamma}_\ell(k)$ (gray box with dashed border) is calculated in the lower path of the diagram. It depends on the third component, the spectral weighting rule (WR), whether the *a posteriori* SNR is required or not. Several spectral weighting rules have been published over the years. For example, one of the first solutions was the well-known Wiener filter (WF), which can be driven easily by an *a priori* SNR estimate only, without requiring the *a posteriori* SNR [McAulay and Malpass, 1980]. Furthermore, the classical minimum mean squared error short-time spectral amplitude (MMSE-STSA) estimator [Ephraim and Malah, 1984] has been improved by minimizing the logarithmic error in [Ephraim and Malah, 1985], thereby yielding the minimum mean squared error log-spectral amplitude (MMSE-LSA) estimator. Both assume a Gaussian distribution of the estimated Fourier coefficients. Later on, the Gaussian assumption has been replaced by a super-Gaussian [Martin, 2002, Martin and Breithaupt, 2003, Martin, 2005]. In [Lotter and Vary, 2005] a super-Gaussian assumption is used together with a maximum a posteriori (MAP) estimator, known as the super-Gaussian joint maximum *a posteriori* (SG-jMAP) estimator. Those rules make use of both, the *a priori* and also the *a posteriori* SNR.

Our goal is the improvement of the *a priori* SNR estimation component which is quite essential for the quality of the enhanced signal. Several methods for *a priori* SNR estimation have been developed over the years such as [Ephraim and Malah, 1984, Cohen, 2005, Plapous et al., 2006, Breithaupt et al., 2008, Suhadi et al., 2011, Nahma et al., 2017, Stahl and Mowlaee, 2018], to name a few. In the following we will provide a short overview on some of the mentioned estimators as they will be used as baselines in the publications for benchmarking.

https://doi.org/10.24355/dbbs.084-202102121510-0

Decision-Directed Estimation

The likely most prominent approach is the frequency domain-based decision-directed (DD) approach as originally published in [Ephraim and Malah, 1984]. The *a priori* SNR calculation is based on the following formula

$$\hat{\xi}_\ell^{\mathrm{DD}}(k) = (1-\beta) \cdot \max\left\{\hat{\gamma}_\ell(k) - 1, 0\right\} + \beta \cdot \frac{|\hat{S}_{\ell-1}(k)|^2}{\hat{\sigma}_{\ell-1}^D(k)^2}. \tag{1.17}$$

Hereby, $\beta \in [0,1]$ is the weighting factor which determines the influence of each estimator w.r.t. their superposition. The first summand is representing a maximum likelihood (ML) *a priori* SNR estimator and the second summand uses the last frame's enhanced speech and noise power estimate to calculate a second estimate. The factor β is often chosen to be close to unity, which causes the DD-based estimate to stay behind by one frame as the influence of the last frame is then very high. This might be an issue, especially in non-stationary environments where a quick responsiveness is beneficial [Cappé, 1994]. Nevertheless, it produces good results and also due to its simplicity it is still commonly used and considered as state of the art.

Harmonic Regeneration

In [Plapous et al., 2006], the authors propose a novel *a priori* SNR estimation approach which is based on the DD approach. They propose it in the context of a speech enhancement system and call it harmonic regeneration noise reduction (HRNR). They compensate the delay of one frame from the DD approach by employing a two-stage noise reduction scheme where the *a priori* SNR estimate is refined in a second stage to explicitly overcome the bias. A time-domain nonlinearity is used, more precisely a half-wave rectifier, to boost and regenerate the harmonic structure of voiced speech frames for *a priori* SNR estimation. This method obtains significant improvement by introducing less harmonic distortion. However, this approach introduces unnatural content prior to the first harmonic [Plapous et al., 2006, Fig. 8], which might lead to artifacts in certain low-frequency noise types. There is no explicit distinction between envelope and excitation with this approach. A precise description of the algorithm to our understanding is presented in Publication **II** [Elshamy et al., 2017a, Sec. II-C 3)].

Cepstro-Temporal Smoothing

Another important *a priori* SNR estimator is the cepstro-temporal smoothing (CTS)-based approach by [Breithaupt et al., 2008]. To our knowledge, it is the first estimator which uses the cepstrum and also discriminates between excitation and envelope. The approach uses

https://doi.org/10.24355/dbbs.084-202102121510-0

the ML *a priori* SNR estimate as a basis to obtain a clean speech power estimate which is subsequently transformed to the cepstral domain by applying an inverse discrete Fourier transform (IDFT). The discrimination is done by addressing the corresponding quefrency bins in the cepstral representation, which is different from discriminating by use of LPC analysis as is done in our proposed approach, even though we also operate in the cepstral domain. The significant difference is the way the separation of excitation and envelope are understood in the cepstral domain and by LPC analysis as explained in Section 1.1. The approach from [Breithaupt et al., 2008] manipulates the coefficients individually w.r.t. to their position in the cepstrum and thus their correspondence to either envelope or excitation. The applied smoothing factors are thus quefrency-dependent and also time-variant. A more detailed introduction to this approach is given in Publication **II** [Elshamy et al., 2017a, Sec. II-C 2)]. This approach has proven to produce high-quality results, especially in non-stationary environments.

https://doi.org/10.24355/dbbs.084-202102121510-0

1.5 Summary of Publication I

The first publication [Elshamy et al., 2015] is basically introducing a prototype system from which we have gained valuable insights, which allowed us to finally end up with the source-filter decomposition in the later publications. However, this first paper does not yet utilize the source-filter model for speech enhancement. Nevertheless, without its provided insights and contributions to our knowledge in this field, the other publications would not have existed in the way they are published now. For this reason, we will briefly summarize its content and also the lessons we have learned during the experiments and the writing of the paper.

The paper is mainly based on the work of [Mowlaee and Saeidi, 2013]. We have reworked the whole math and reformulated it in a more concise way according to our understanding and thereby simplified the reproducibility and also the comprehensibility. The general idea was to train a Gaussian mixture model (GMM) with clean speech spectral prototypes, which is then used to find the corresponding clean spectrum for a noisy, or in this case preliminary denoised, observation. Preliminary denoised, because a traditional first denoising stage was applied to obtain a more suitable spectral representation to match the GMM against. Originally, this intermediate clean speech estimate, together with a noise power estimate, defined the spectral weighting rule, which has been used to retrieve the final clean speech estimate from the microphone signal. However, we found that the use of the clean speech estimate as the numerator of an *a priori* SNR estimate yields several advantages: We could broaden the field of application significantly and found the system to be more robust against false estimations of the GMM. In addition to that, we introduced an iterative scheme which continuously re-evaluated the GMM until convergence is reached, leading to further improvement of the method. However, we observed that the system was incapable of removing noise between the harmonics of speech. Our analyses have shown, that during the training of the GMM most of the harmonic excitation components of the spectral speech prototypes must have been lost due to averaging processes. Furthermore, the few remaining modes, which would represent harmonic components, are selected only in rare cases, as a very precise match of fundamental frequency between input signal and trained speech spectral prototypes would be required for a good match. This led us directly to the idea of treating spectral envelope and excitation signal separately, as it seemed that too much information, as comprised in a clean speech spectral prototype, could not be represented with sufficient precision by a single GMM. Thus, the use of the source-filter model was obvious. Those findings have carried, motivated, and influenced the remaining publications substantially. However, this in retrospective quite crude method has worked surprisingly well and exceeded the performance of the DD state-of-the-art estimator by obtaining substantially higher noise attenuation results for low-SNR conditions in combi-

https://doi.org/10.24355/dbbs.084-202102121510-0

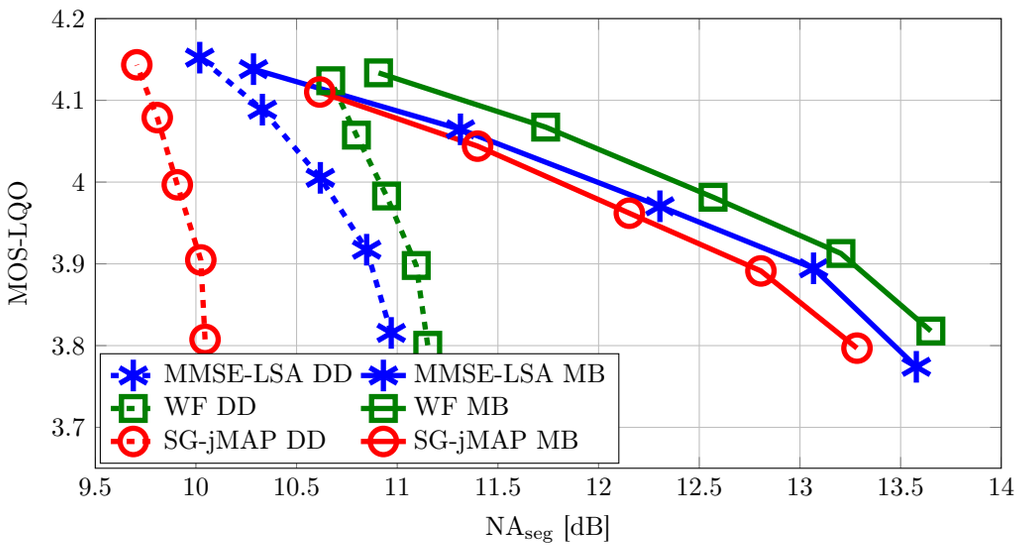


Figure 1.5: Main results of Publication **I** showing the performance of the proposed model-based (MB) *a priori* SNR estimator (solid lines), compared to the DD baseline (dashed lines) in combination with three different spectral WRs, distinguishable by the markers. Each marker represents a specific SNR condition ranging from -5 dB (bottom) to 15 dB (top) in steps of 5 dB.

nation with various spectral weighting rules. The main results of the publication are shown in Figure 1.5, where the MOS-LQO is plotted over the segmental NA with each marker representing a certain SNR condition ranging from -5 dB to 15 dB in steps of 5 dB (bottom to top). The advantage in the low-SNR conditions (lower markers) is obvious due to the strong right shift of the plots in such low SNRs.

https://doi.org/10.24355/dbbs.084-202102121510-0

2 Enhancing the Excitation Signal

This chapter deals with the major contributions of this thesis in the field of excitation signal enhancement in the context of *a priori* SNR estimation for speech enhancement. A brief overview over state-of-the-art approaches is given, however, this field has not been researched very extensively yet and it is more a general overview of excitation enhancement methods. We present our advances with traditional signal processing-based methods, which are subsequently substituted by more modern approaches based on machine learning in Section 2.2. A short conclusion is given in Section 2.3 to sum up our findings and the most important contributions in that field.

2.1 State of the Art

To our knowledge, there are only few approaches which deal with the enhancement of the excitation signal obtained by any given model to separate speech into source and filter. The most relevant approach is the CTS method by [Breithaupt et al., 2008], as it is also used for *a priori* SNR estimation, which makes it perfectly comparable. However, even though the approach operates in the cepstral domain as ours does, it does not follow the separation by an explicit LPC model but uses the cepstral liftering method. The coefficients that are supposed to represent the excitation signal are smoothed in a special way compared to the remaining coefficients of the cepstrum. The separation in the cepstral domain is also not guaranteed to deliver consistent results, e.g., as the liftering might affect the pitch if the cut-off quefrency is not chosen wisely and thus, the dominant component of the excitation. The excitation signal is not necessarily "flat" in the sense that its envelope is more or less spectrally flat as provided by LPC analysis. Also, the enhancement is of a different nature compared to the proposed method as it does not explicitly address or enhance specific properties of the excitation but only applies a smoothing to reduce noise. Due to these facts an improvement by using LPC analysis and to directly manipulate the excitation signal seems reasonable and promising.

Despite this, there are some further publications that should be mentioned, even though they might not be directly comparable, but for the sake of completeness. In [Yegnanarayana and Murthy, 2000] the authors propose the enhancement of an LPC residual signal for speech dereverberation. Here, a dynamic weighting function is derived and used to enhance

https://doi.org/10.24355/dbbs.084-202102121510-0

or manipulate the residual signal in three different regions to reduce the reverberation. After manipulation, the residual signal is used to excite the all-pole filter to synthesize the enhanced speech signal. However, due to its focus on dereverberation and the synthesis of the enhanced signal, it is rather unsuitable for comparison. Excitation source information is used in [Gandhi et al., 2006] to calculate a weighting function and to enhance a residual signal obtained by LPC analysis with it. Interestingly, the authors also use a first noise reduction stage, here spectral subtraction, to obtain a signal which is then further manipulated. Finally, the enhanced signal is synthesized with the modified excitation signal and contains less musical noise. Due to its nature as a postprocessor for the spectral subtraction method and the fact that the enhanced signal is synthesized, which has to be done very carefully in speech enhancement, it is also unsuitable for direct comparison. In [Simsekli et al., 2014], the authors propose a novel approach for noise reduction, where the source-filter model is used and the excitation and spectral envelope are modeled as non-negative dynamic systems, as the authors would call it. Interestingly, the authors also make use of a preliminary denoising stage to obtain a better suited signal for further processing. However, it is a quite complex algorithm that makes use of a separate pitch estimation algorithm [Talkin, 1995], which indicates that it might be inadequate for the use in telephony applications.

2.2 Summary of Publications II, III, VI, and VII

In the following, we will briefly summarize Publications **II**, **III**, **VI**, and **VII**, which document our progress in the field of excitation signal enhancement, which is the first thematic complex of this thesis. The publication scheme is that a comprehensive journal publication is followed by a condensed conference extract, which provides some additional analysis, variant or evaluation. The first two publications present a traditional signal processing-based solution while the subsequent publications investigate the benefits of a neural network in that context and also provide results of a subjective listening test. Please note that in the temporal course two further Publications **IV** and **V** have been written in between, where the enhancement of the spectral envelope is addressed. In that context some variants have been proposed that use excitation and envelope enhancement jointly which will be mentioned in this chapter briefly. Those variants are explained in further detail later on in the respective summaries in Section 3.2. We have decided for that order to provide a more coherent structure in terms of thematic complexes.

https://doi.org/10.24355/dbbs.084-202102121510-0

Publication II

Publication **II** [Elshamy et al., 2017a] deals exclusively with the manipulation of the excitation signal for instantaneous *a priori* SNR estimation in the context of a speech enhancement system. We introduced a classical signal processing-based scheme in the cepstral domain to reinforce the harmonic structure of the excitation signal. The so-called cepstral excitation manipulation (CEM) is a method that is natively working for voiced and unvoiced frames without requiring a dedicated algorithm for their discrimination. After a preliminary noise reduction stage, which was motivated by the findings from Publication **I** and subsequent LPC analysis to obtain the excitation signal, the properties of the cepstral domain are exploited to synthesize an artificial idealized excitation signal which is very low in complexity.

Two different methods are investigated, where the first synthesizes a pure cosine-like excitation signal and the second uses a storage of pre-trained excitation templates to replace most of the observed excitation signal with. In addition to that, speaker-dependent and speaker-independent options are tested. All algorithms depend on a simple pitch estimation which is also operating in the cepstral domain and the system has proven to be surprisingly robust against estimation errors. The algorithms are enriched with a simple mechanism that adds a more natural transient to the first and last harmonic's rising and falling edge, respectively. This is due to the fact that investigations have revealed that this is naturally occuring within some of the pre-trained excitation templates, but not for all of them and not for the synthesized cosine-like excitation signals. In Figure 2.1, one can see an example of a stored excitation template in the upper panel, where the smooth transition of the first and last harmonic can be seen. The lower panel depicts a synthesized excitation with applied start and end decay of the first and last harmonic, respectively, which nicely mimics the natural course as seen in the upper panel. It is also to be seen how the synthesized harmonics coincide nicely with the corresponding template's harmonics. To our knowledge, the system is quite robust and works well for unvoiced excitation signals, too. This is because in an unvoiced excitation signal's cepstrum, the pitch estimator will not identify a distinct high-amplitude quefrency bin, but rather one that is low in amplitude and comparable to the others. Consequently, the reinforcement, which is based on this identified quefrency bin's amplitude, is not likely to produce a prominent harmonic structure. The manipulated excitation signal is mixed with the spectral envelope of the preliminary denoised signal, i.e., the signal after the first noise reduction stage, to finally yield the numerator for an instantaneous *a priori* SNR estimate. The instantaneous fashion yields a more responsive estimator that is independent of previous frames as compared to the DD approach.

The approaches are evaluated in a comprehensive way by using four measures. To assess the noise attenuation performance, the segmental NA as well as the delta SNR are evaluated.

https://doi.org/10.24355/dbbs.084-202102121510-0

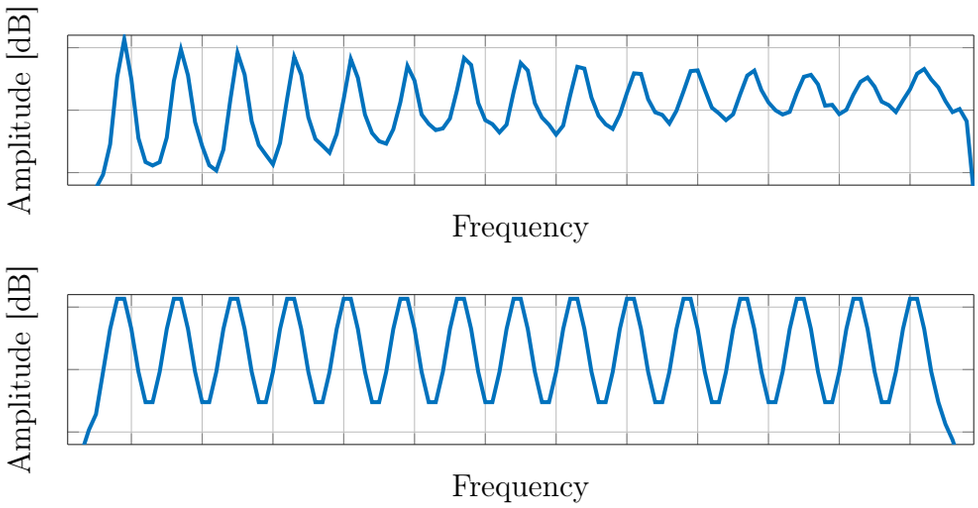


Figure 2.1: Upper panel: Example of a log-spectrum excitation template. Lower panel: Example of an idealized synthetic excitation log-spectrum with applied start and end decay.

For the speech component quality the segmental SSDR and the MOS-LQO are considered. Two extensive tables depict the detailed results for several SNR conditions and four different noise types, among them two non-stationary noises. The in total three proposed approaches are evaluated against three baselines namely: the DD, HRNR, and CTS approaches. To broaden the analysis, two different spectral WRs haven been used, MMSE-LSA and SG-jMAP. The three proposed methods provide significantly higher delta SNR scores and slightly higher segmental NA while maintaining comparable speech component quality. The template-based approach obtained slightly better performance compared to the synthesized alternative. However, going further and introducing speaker-dependent templates did not yield any significant advantage. In total, we could introduce a very low-complex and simple approach that outperformed several state-of-the-art approaches. Furthermore, a patent has been granted in several countries for the CEM technique [Elshamy et al., 2017b, Elshamy et al., 2019b, Elshamy et al., 2019a] and is likely to be found in some variant in a number smartphones.

Publication III

Publication **III** [Elshamy et al., 2017c] is a distilled version of the corresponding journal Publication **II**. However, it presents a generic variant of the earlier proposed CEM approaches, where the speaker-independent and template-based CEM method is not introduced as stand-alone *a priori SNR estimator* but in the context of a two-stage *speech*

https://doi.org/10.24355/dbbs.084-202102121510-0

Table 2.1: Results of Publication **III** showing the summarizing results of the figure of merit for the compared approaches for six SNRs and four noise types. The best performing approach is in **boldface**.

		FoM						
	SNR [dB]	-5	0	5	10	15	20	mean
ROAD	MMSE-LSA	0.86	0.92	0.97	1.02	1.05	0.98	0.97
	SG-jMAP	0.90	0.95	1.01	1.06	1.09	1.02	1.01
	HRNR	0.88	0.93	0.98	1.02	1.05	0.98	0.97
	MMSE-LSA CEM	0.92	0.98	1.03	1.06	1.08	1.03	1.02
	SG-jMAP CEM	**0.93**	**0.99**	**1.05**	**1.09**	**1.11**	**1.04**	**1.03**
CAR	MMSE-LSA	0.90	0.95	0.97	0.99	1.00	0.97	0.96
	SG-jMAP	0.92	0.96	0.99	1.01	1.01	0.98	0.98
	HRNR	0.92	0.94	0.93	0.91	0.89	0.91	0.92
	MMSE-LSA CEM	**1.04**	**1.08**	**1.10**	**1.11**	**1.10**	**1.09**	**1.09**
	SG-jMAP CEM	1.00	1.04	1.07	1.08	1.07	1.05	1.05
OFFICE	MMSE-LSA	0.78	0.89	0.98	1.04	1.09	0.98	0.96
	SG-jMAP	0.80	0.91	0.99	1.06	1.10	1.00	0.98
	HRNR	0.74	0.84	0.92	0.97	1.01	0.92	0.90
	MMSE-LSA CEM	**0.92**	**1.05**	**1.14**	**1.20**	**1.23**	**1.13**	**1.11**
	SG-jMAP CEM	0.86	0.99	1.08	1.14	1.17	1.07	1.05
PUB	MMSE-LSA	0.74	0.89	1.01	1.09	1.16	1.02	0.98
	SG-jMAP	0.74	0.90	1.02	1.11	1.18	1.03	0.99
	HRNR	0.67	0.83	0.94	1.03	1.10	0.95	0.92
	MMSE-LSA CEM	**0.75**	**0.97**	**1.10**	**1.19**	**1.25**	**1.09**	**1.06**
	SG-jMAP CEM	**0.75**	0.95	1.08	1.17	1.24	1.08	1.05
Means	MMSE-LSA	0.82	0.91	0.98	1.04	1.07	0.99	0.97
	SG-jMAP	0.84	0.93	1.00	1.06	1.10	1.01	0.99
	HRNR	0.80	0.88	0.94	0.98	1.01	0.94	0.93
	MMSE-LSA CEM	**0.91**	**1.02**	**1.09**	**1.14**	**1.17**	**1.08**	**1.07**
	SG-jMAP CEM	0.88	0.99	1.07	1.12	1.15	1.06	1.05

enhancement system. It depicts a good way the algorithm would actually be used in practice. Even though it is not required to be used solely as stand-alone *a priori* SNR estimator, it was perfectly reasonable to do so in Publication **II** for reasons of comparability. The two-stage speech enhancement system is even less complex than the stand-alone estimator. It

https://doi.org/10.24355/dbbs.084-202102121510-0

is also evaluated against a traditional system with the DD *a priori* SNR estimator and also against the HRNR approach. The DD and CEM approaches are evaluated with the MMSE-LSA and the SG-jMAP spectral WRs under the same four objective metrics as in Publication **II**. Furthermore, as the interpretation and evaluation of the many measures for six SNR conditions and four noise types might be cumbersome, we additionally introduced a figure of merit (FoM) which combines the four measures to one single score to allow for an easier evaluation. The FoM considers each measure equally and is defined as follows:

$$\mathrm{FoM} = \frac{1}{4}\frac{\mathrm{NA_{seg}}}{\overline{\mathrm{NA_{seg}}}} + \frac{1}{4}\frac{\Delta\mathrm{SNR}}{\overline{\Delta\mathrm{SNR}}} + \frac{1}{4}\frac{\mathrm{MOS\text{-}LQO}}{\overline{\mathrm{MOS\text{-}LQO}}} + \frac{1}{4}\frac{\mathrm{SSDR_{seg}}}{\overline{\mathrm{SSDR_{seg}}}}. \tag{2.1}$$

Here, the normalizing denominators represent the average of the respective measure over all tested SNR conditions and noise types.

The final results of Publication **III** are shown in Table 2.1, where the best results are in boldface type. It can be seen that the proposed CEM approach always outperforms its respective baseline and thus represents the superior method under the FoM. The approach obtains significantly higher delta SNR and segmental NA, while maintaining a similar speech component quality, when compared to the baselines and thus can be seen as more balanced.

Publication VI

In Publication **VI** [Elshamy and Fingscheidt, 2019a], the originally published traditional signal processing-based approach from Publication **II** is enhanced by introducing a deep neural network (DNN) for CEM. The enhanced CEM approach is used in the same framework for *a priori* SNR estimation as before. Several aspects of the training process for the neural network are analyzed and optimized, among them we investigated effects of input feature normalization, target normalization, different targets as well as various topologies w.r.t. number of layers and nodes. As the DNN is supposed to enhance cepstral excitation signals, the intuitive way is to also extract the training targets from clean speech signals. Still, in order to reconstruct a perfect clean speech signal, this implicitly assumes that the clean speech spectral envelope is available, which is not the case. An alternative is to consider also the spectral envelope of the observed signal and trying to incorporate that information into the DNN by using training targets that are obtained by using the LPC coefficients of the actually observed signal. An oracle experiment has shown improved results for that kind of targets compared to using clean speech excitation signals as targets. However, in a practical system the required statistics, i.e., mean and variance, of the better suited targets would not be available. The statistics are required for the rescaling of the DNN's output and for this kind of targets it is only possible to obtain them in lab conditions. This is a problem since we could also show that the normalization of the targets is

https://doi.org/10.24355/dbbs.084-202102121510-0

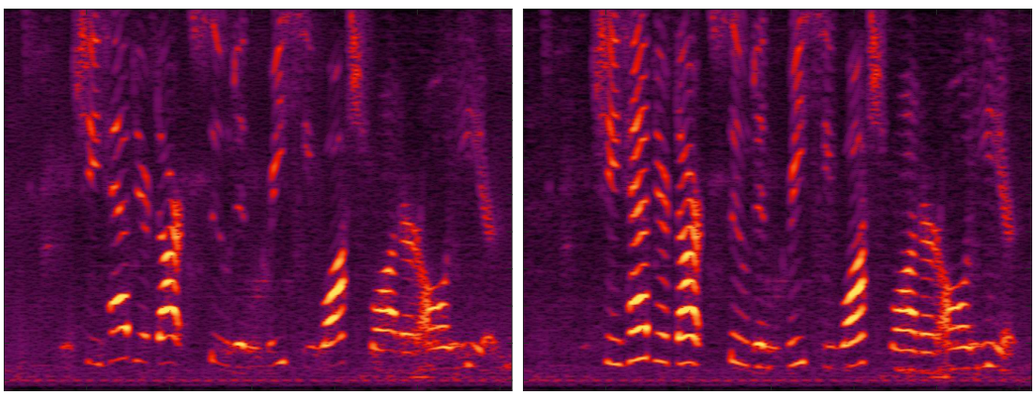


Figure 2.2: Spectrograms of an enhanced microphone signal at 10 dB SNR with non-stationary noise processed by DNN-based CEM without (left) and with mean/variance normalization (right) of the targets.

a key to success as can be seen in Figure 2.2. Here, the left image depicts the spectrogram of an enhanced signal by aid of a DNN-based CEM approach without target normalization and the right image nicely shows the benefits of applied target normalization as the harmonic structure is preserved much better and the spectrogram thus turns out to be much richer.

The approach is evaluated against several *a priori* SNR estimator baselines and a modern ideal ratio mask-based speech enhancement approach under various objective measures such as the segmental NA, delta SNR to assess the NA performance and for the speech component quality and intelligibility, MOS-LQO and STOI are evaluated, respectively.

Please note that the results cannot be directly compared to the results from Publication **II**. Due to the investigation of speaker-dependent and speaker-independent excitation templates in Publication **II** a different database setup has been used. However, for valid training and evaluation of DNNs three disjoint datasets are required for training, development, and testing. Thus more data was required and the use of multiple databases for speech and noise became necessary. However, the results for the CEM baseline on the new data have been reported in Publication **VI**, too.

The new approach shows a strong improvement in terms of NA, while being able to compete easily in terms of speech component quality and also intelligibility.

https://doi.org/10.24355/dbbs.084-202102121510-0

Publication VII

Publication **VII** [Elshamy and Fingscheidt, 2019b] is a condensed version of the corresponding journal Publication **VI**. The framework for *a priori* SNR estimation is used in the same setting with a DNN-based solution for CEM. However, another variant is introduced, which respects the inherent energy coefficient c_0 from the observed excitation signal, instead of replacing it with the predicted value delivered by the DNN. This minor modification leads to an improvement of the speech component quality and also slightly increases intelligibility at the cost of less NA performance, still outperforming the traditional signal processing-based CEM approach in both dimensions, i.e., an absolute overall improvement.

A major contribution of this paper is a semi-formal comparison category rating (CCR) subjective listening test, which was able to verify the superiority of our proposed approach as already shown by objective measures. The results report the comparison mean opinion score (CMOS), which shows a significant preference of the listeners for our proposed approach over the traditional DD *a priori* SNR estimation baseline. In total, 17 non-professional subjects participated in the listening test, where two of the subjects always preferred the noisy condition over the processed conditions. Here, we do not know if this was intention or misunderstanding. For that reason we show two tables with results considering all the 17 subjects in Table 2.2 and results from only 15 subjects without the two potential outliers in Table 2.3. All participating subjects were required to be native speakers w.r.t. the used language for the samples of the listening test.

Please note that "$\rightarrow$" indicates the serial concatenation of the spectral envelope enhancement method called cepstral envelope enhancement (CEE) and excitation enhancement (CEM-DNN/CEM-DNN-c_0) in either order. This serial concatenation is explained in more detail in Publications **IV** and **V** which are summarized in Section 3 due to the thematic grouping in this thesis and the otherwise different chronological sequence of publication.

The results show that in either case with or without outliers, a strong significant preference of our method CEE $\rightarrow$ CEM-DNN was reported for condition d) when tested against the DD baseline. However, a further strong preference among the proposed variants could not be reported. The variants where rated very similar among the subjects as can be seen for conditions e) and f) where CEM-DNN-c_0 denotes the aforementioned variant that does not predict the energy coefficient. The significance results from the fact that the null hypothesis, meaning that the respective approach is not preferred over the other (CMOS = 0), is not included in the reported confidence interval.

https://doi.org/10.24355/dbbs.084-202102121510-0

Table 2.2: CMOS results and 95 % confidence intervals for the subjective listening test. The preferred approach is in **boldface**.

Condition	CMOS	CI_{95}
a) Noisy vs. **DD**	0.96	± 0.16
b) Noisy vs. **CEM-DNN**	0.97	± 0.22
c) Noisy vs. **CEE → CEM-DNN**	1.22	± 0.20
d) DD vs. **CEE → CEM-DNN**	**0.25**	± 0.17
e) CEM-DNN vs. **CEE → CEM-DNN**	0.12	± 0.11
f) CEE → CEM-DNN-c_0 vs. **CEE → CEM-DNN**	0.07	± 0.13

Table 2.3: CMOS results and 95 % confidence intervals for the subjective listening test with two outlier subjects not considered. The preferred approach is in **boldface**.

Condition	CMOS	CI_{95}
a) Noisy vs. **DD**	1.19	± 0.16
b) Noisy vs. **CEM-DNN**	1.35	± 0.19
c) Noisy vs. **CEE → CEM-DNN**	1.62	± 0.16
d) DD vs. **CEE → CEM-DNN**	**0.48**	± 0.16
e) CEM-DNN vs. **CEE → CEM-DNN**	0.15	± 0.11
f) CEE → CEM-DNN-c_0 vs. **CEE → CEM-DNN**	0.17	± 0.14

2.3 Conclusion

We have proposed a generic method for the enhancement of excitation signals in the context of an instantaneous *a priori* SNR estimator for speech enhancement. However, the method is not restricted to that specific application and can theoretically be used for every other algorithm that requires an *a priori* SNR estimate or deals with the enhancement of excitation signals. The combination of using LPC analysis to obtain the source and the filter together with the introduction of a cepstral processing scheme has not been investigated before and is a key to the success of the proposed CEM approaches. We have first introduced a simple traditional signal processing-based solution, which has been developed further to integrate a DNN into the enhancement process. For this we could show the importance of input feature and target normalization for a successful application. The results show improved preservation of harmonic structures and a richer spectrum compared to baseline approaches. Due to the instantaneous estimation, there is no delay in the responsiveness of the estimator as compared to the DD approach.

https://doi.org/10.24355/dbbs.084-202102121510-0

The traditional signal processing-based CEM approach could achieve a delta SNR improvement of more than 2 dB while maintaining a comparable speech component quality when compared to the DD approach. The DNN-based CEM solution could contribute to an absolute improvement of up to 1.5 dB segmental NA over traditional CEM and up to 3 dB segmental NA over the DD approach without any significant degradation of the speech component. Finally, the DNN-based CEM solution in concatenation with the spectral envelope enhancement method CEE led to further improvement of up to 2 dB segmental NA over traditional CEM and up to 3.5 dB segmental NA over the DD approach. Furthermore, a semi-formal subjective CCR listening test has shown that the listeners prefer the proposed DNN-based CEM in concatenation with the CEE approach over DD by 0.25 CMOS points or even by 0.48 CMOS points when two outlier subjects are not considered. The CEM technique has been filed as a patent in Europe, the United States, and China and as of today, has already been granted in Europe and in the United States.

https://doi.org/10.24355/dbbs.084-202102121510-0

3 Enhancing the Spectral Envelope

This chapter deals with the major contributions of this thesis in the field of spectral envelope enhancement in the context of *a priori* SNR estimation for speech enhancement. We will provide an overview over state-of-the-art approaches in the following section, focusing on applications for noise reduction purposes, however several speech signal processing methods work with the spectral envelope, also in different fields such as automatic speech recognition (ASR) or artificial bandwidth extension (ABE), which partly has inspired this work as well. Our advances in the field of spectral envelope enhancement for noise reduction are reported in Section 3.2, where traditional models and also modern DNN-based solutions are evaluated. Finally, the most important results and findings are summed up in a conclusion in Section 3.3.

3.1 State of the Art

Several approaches address various speech enhancement problems by enhancing the spectral envelope. For example, in [Srinivasan and Samuelsson, 2003, Srinivasan et al., 2006, Srinivasan et al., 2007] the authors propose a noise reduction algorithm based on two codebooks with spectral envelope prototypes for speech and noise. The correct envelopes or codebook entries are inferred by finding a combination of prototypes and gain factors that matches the observation. The codebook entries are then used in a WF together with their respective gain factors to retrieve the enhanced speech signal. The approach has been brought to the cepstral domain in [Rosenkranz, 2010]. However, using only spectral envelopes prevents modeling of the fine structure that is representing the excitation signal, which is a major drawback. As already stated in Section 2.1, the approach in [Simsekli et al., 2014] is also using the source-filter model and therein the spectral envelope is modeled as non-negative dynamic system, as the authors would call it. However, for the above-mentioned reason it seems unsuitable for telephony applications. A first approach, that utilizes a hidden Markov model (HMM), was proposed in [Ephraim, 1992] and was developed further in [Zhao and Kleijn, 2007]. However, both use a low order model to represent speech and noise as autoregressive processes which suggests that they suffer from the same problem of being incapable of modeling the fine structure.

HMMs are also used for ASR [Furui, 2000] and ABE [Jax and Vary, 2007]. They are

https://doi.org/10.24355/dbbs.084-202102121510-0

frequently used to find a mapping of an observation and the corresponding hidden state, which is responsible for the observation. Using GMMs as a backend was commonly done until DNNs came into play. Hinton et al. introduced DNNs as novel backend for HMMs for ASR with great success in [Hinton et al., 2012]. A similar development has been observed for ABE, where in [Abel et al., 2016, Abel and Fingscheidt, 2017, Abel and Fingscheidt, 2018] the transition from HMM with GMM backend over HMM with DNN backend to finally DNN-only solutions is nicely depicted.

The most relevant prior art in the specific field of spectral envelope enhancement for *a priori* SNR estimation is also the CTS approach [Breithaupt et al., 2008]. It addresses envelope and excitation signal separately with specific smoothing factors in the cepstral domain and uses the results directly for *a priori* SNR estimation. However, the lack of a specific model with constraints as enforced by LPC analysis leaves room for improvement. We could show already in Publication **II**, that we are able to outperform the CTS approach by only addressing and manipulating the excitation signal explicitly. This rendered a further comparison against CTS needless and allowed us to use our own approach as a baseline together with the DD approach as an anchor. Most naturally, this led to the investigation of a scheme combining our separate enhancement methods of excitation and envelope.

3.2 Summary of Publications IV and V

The following summaries of Publications **IV** and **V** will document our progress in the field of envelope enhancement. Analogous to Section 2.2, a comprehensive journal publication is followed by a shorter conference publication that provides further insight. Both represent the second thematic complex, however, they are chronologically located between Publications **III** and **VI**, when considering the development process of the entire system.

Publication IV

In Publication **IV** [Elshamy et al., 2018b], we investigate various methods for the enhancement of spectral envelopes in the cepstral domain, dubbed cepstral envelope enhancement (CEE), for *a priori* SNR estimation in the context of speech enhancement. Based on noisy observations, we investigate the ability of an HMM driven by GMMs as acoustic model to infer the corresponding clean spectral envelope. The GMM-based system is subsequently enhanced with DNNs, which replace the former acoustic model as it has been done before in various other contexts and shown great improvement. Consequently, the codebook-based backend of the HMMs is used with either MMSE or MAP estimation to obtain the final estimate, where MMSE clearly performs better. This is likely due to its ability to use the whole codebook space instead of a single entry. The system is then evolved to a solely

https://doi.org/10.24355/dbbs.084-202102121510-0

DNN-based solution either for classification in combination with the codebook or regression, which then directly estimates the envelope representation in the cepstral domain. Several parameters for the DNN training have been optimized such as the number of layers and nodes or the type of activation functions. Due to the data requirements for the training process we had to use multiple speech and noise databases. Furthermore, observing a different distribution of speech-active and speech-inactive segments across the databases, it was necessary to conduct an investigation that discriminates between those two classes to get the full picture of the performance on the more important speech active frames.

In addition to that, advances towards a system that uses excitation and envelope enhancement jointly were made to tap the full potential of both schemes. Therefore, either a serial or a parallel structure of CEM and CEE is used where the serial concatenation of first applying CEE followed by CEM has proven to deliver best results. The results have been evaluated using the segmental NA and the delta SNR to assess the noise attenuation performance, and MOS-LQO and STOI to assess the speech component quality and intelligibility, respectively. Final independent results are reported on a test set with four unseen noise types containing stationary and non-stationary noises.

The systems are benchmarked against the traditional DD *a priori* SNR estimator, the former introduced CEM approach from Publication **II**, and a modern ideal ratio mask-based approach. The performance of the proposed stand-alone CEE approach is already quite convincing, however, in serial concatenation with CEM the performance could be further improved by increasing the noise attenuation performance while preserving speech component quality and speech intelligibility on a comparable level. It is recommended to use a serial scheme employing CEE first and CEM second.

Publication V

Publication **V** [Elshamy et al., 2018a] is the distilled conference contribution of the corresponding journal Publication **IV**. It picks up some of the results from Publication **IV** and extends the evaluation by using two additional spectral WRs, namely the WF and also the SG-jMAP. This provides a larger picture of the performance especially in a speech enhancement context. Due to space limitations the evaluation has been restricted to delta SNR and segmental SSDR. The performance of the proposed approaches compared to the baselines can be seen in Figure 3.1. The results support the recommendation of using a serial concatenation of CEE followed by CEM and also show that for each spectral WR the results are improved, which indicates indirectly that the *a priori* SNR is actually improved.

https://doi.org/10.24355/dbbs.084-202102121510-0

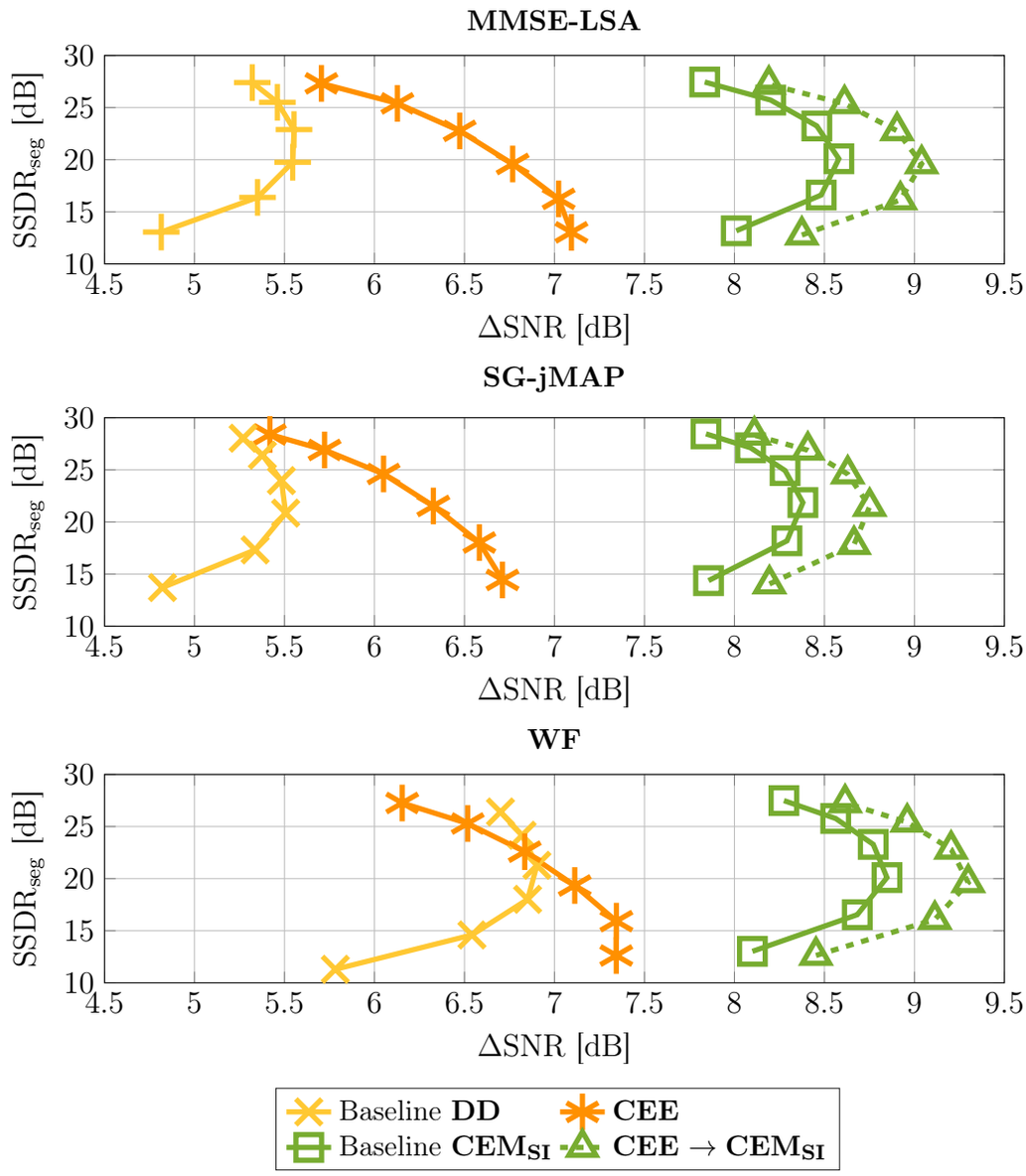


Figure 3.1: Main results of Publication **V** showing the evaluation of SSDR_{seg} and ΔSNR for the *a priori* SNR estimators under test in non-stationary and unseen noises, together with three different spectral weighting rules. Each marker represents a specific SNR condition ranging from -5 dB (bottom) to 20 dB (top) in steps of 5 dB.

https://doi.org/10.24355/dbbs.084-202102121510-0

3.3 Conclusion

We have investigated various approaches in the cepstral domain for the enhancement of spectral envelopes for *a priori* SNR estimation in the context of a noise reduction algorithm. We do not want to restrict the application to that specific use case; however, further applications would have to be investigated. Based on the findings in Section 2.2, we continued to use LPC analysis to separate a preliminary denoised signal into the excitation signal and the spectral envelope. The latter is then used for the proposed CEE methods to estimate a clean spectral envelope by means of an HMM, first driven by GMMs and later on by a DNN. We investigated MAP and MMSE estimation, which has shown that MMSE is the superior method. Subsequently, the whole HMM was replaced by a DNN. The superiority of a classification DNN together with a codebook could be shown in several experiments, also for non-stationary and unseen noises. A key finding of this work was the successful serial combination of CEE and CEM. This allows a joint enhancement of excitation and envelope which resulted in an improved NA performance by 0.5 dB over the earlier proposed CEM approach and even by 2 dB over the traditional DD approach, while the speech component quality and intelligibility remains quite comparable. We could also show a consistent improvement for three spectral WRs which indicates an actual improvement of the *a priori* SNR.

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

Bibliography

[Abel and Fingscheidt, 2017] J. Abel and T. Fingscheidt, "A DNN Regression Approach to Speech Enhancement by Artificial Bandwidth Extension," in *Proc. of WASPAA*, New Paltz, NY, USA, Dec. 2017, pp. 219–223.

[Abel and Fingscheidt, 2018] J. Abel and T. Fingscheidt, "Artificial Speech Bandwidth Extension Using Deep Neural Networks for Wideband Spectral Envelope Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, Jan. 2018.

[Abel et al., 2016] J. Abel, M. Strake, and T. Fingscheidt, "Artificial Bandwidth Extension Using Deep Neural Networks for Spectral Envelope Estimation," in *Proc. of IWAENC*, Xi'an, China, Sep. 2016, pp. 1–5.

[Benesty et al., 2008] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin: Springer, 2008.

[Breithaupt et al., 2008] C. Breithaupt, T. Gerkmann, and R. Martin, "A Novel A Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing," in *Proc. of ICASSP*, Las Vegas, NV, USA, Mar. 2008, pp. 4897–4900.

[Cappé, 1994] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[Cohen, 2003] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[Cohen, 2005] I. Cohen, "Speech Enhancement Using Super-Gaussian Speech Models and Noncausal A Priori SNR Estimation," *Speech Communication*, vol. 47, no. 3, pp. 336–350, Nov. 2005.

[Elshamy and Fingscheidt, 2019a] S. Elshamy and T. Fingscheidt, "DNN-Based Cepstral Excitation Manipulation for Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1803–1814, Nov. 2019.

https://doi.org/10.24355/dbbs.084-202102121510-0

[Elshamy and Fingscheidt, 2019b] S. Elshamy and T. Fingscheidt, "Improvement of Speech Residuals for Speech Enhancement," in *Proc. of WASPAA*, New Paltz, NY, USA, Oct. 2019, pp. 214–218.

[Elshamy et al., 2015] S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "An Iterative Speech Model-Based A Priori SNR Estimator," in *Proc. of Interspeech*, Dresden, Germany, Sep. 2015, pp. 1740–1744.

[Elshamy et al., 2017a] S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "Instantaneous A Priori SNR Estimation by Cepstral Excitation Manipulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[Elshamy et al., 2017b] S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "SIGNAL PROCESSOR," Chinese Patent CN107 437 421A, Dec. 05, 2017.

[Elshamy et al., 2017c] S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "Two-Stage Speech Enhancement with Manipulation of the Cepstral Excitation," in *Proc. of HSCMA*, San Francisco, CA, USA, Mar. 2017, pp. 106–110.

[Elshamy et al., 2018a] S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "A Priori SNR Computation for Speech Enhancement Based on Cepstral Envelope Estimation," in *Proc. of IWAENC*, Tokyo, Japan, Sep. 2018, pp. 531–535.

[Elshamy et al., 2018b] S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "DNN-Supported Speech Enhancement With Cepstral Estimation of Both Excitation and Envelope," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2460–2474, Dec. 2018.

[Elshamy et al., 2019a] S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "A SIGNAL PROCESSOR," European Patent EP3 242 295B1, Sept. 20, 2019.

[Elshamy et al., 2019b] S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "SIGNAL PROCESSOR," U.S. Patent US10 297 272B2, May 21, 2019.

[Ephraim, 1992] Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[Ephraim and Malah, 1984] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[Ephraim and Malah, 1985] Y. Ephraim and D. Malah, "Speech Enhancement Using a Min-

https://doi.org/10.24355/dbbs.084-202102121510-0

imum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[Fingscheidt et al., 2008] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-Optimized Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, May 2008.

[Flanagan, 1965] J. L. Flanagan, *Speech Analysis Synthesis and Perception.* Berlin, Germany: Springer, 1965.

[Furui, 2000] S. Furui, *Digital Speech Processing, Synthesis, and Recognition, Second Edition.* Boca Raton, FL, USA: Taylor & Francis, 2000.

[Gandhi et al., 2006] N. Gandhi, P. Krishnamoorthy, and S. R. M. Prasanna, "Reduction of Musical Noise in Spectral Subtracted Speech Using Excitation Source Information," in *Proc. of Annual IEEE India Conference*, New Delhi, India, Sep. 2006.

[Gerkmann and Hendriks, 2012] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[Gustafsson et al., 1996] S. Gustafsson, R. Martin, and P. Vary, "On the Optimization of Speech Enhancement Systems Using Instrumental Measures," in *Proc. of Workshop on Quality Assessment in Speech, Audio, and Image Communication*, Darmstadt, Germany, Mar. 1996, pp. 36–40.

[He et al., 2017] Q. He, F. Bao, and C. Bao, "Multiplicative Update of Auto-Regressive Gains for Codebook-Based Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 457–468, Mar. 2017.

[Hendriks et al., 2010] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE Based Noise PSD Tracking With Low Complexity," in *Proc. of ICASSP*, Dallas, TX, USA, 2010, pp. 4266–4269.

[Hinton et al., 2012] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," in *IEEE Signal Processing Magazine*, vol. 29, no. 6, Nov. 2012, pp. 82–97.

[Hu and Loizou, 2008] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, jan 2008.

https://doi.org/10.24355/dbbs.084-202102121510-0

[Huang et al., 2018] Q. Huang, C. Bao, and X. W. a. Yang Xiang, "DNN-Based Speech Enhancement Using MBE Model," in *Proc. of IWAENC*, Tokyo, Japan, Sep. 2018, pp. 196–200.

[ITU, 1996] ITU, *Rec. P.800: Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Aug. 1996.

[ITU, 2001] ITU, *Rec. P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Feb. 2001.

[ITU, 2003] ITU, *Rec. P.862.1: Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Nov. 2003.

[ITU, 2007] ITU, *Rec. P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Nov. 2007.

[ITU, 2011] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.

[Jax and Vary, 2007] P. Jax and P. Vary, "Wideband Extension of Telephone Speech Using a Hidden Markov Model," in *Proc. of IEEE Workshop on Speech Coding*, Delavan, WI, USA, Sep. 2007, pp. 133–135.

[Lotter and Vary, 2005] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, May 2005.

[Markel and Gray, 1976] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. Berlin, Germany: Springer, 1976.

[Martin, 2001] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[Martin, 2002] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors," in *Proc. of ICASSP*, Orlando, FL, USA, May 2002, pp. 253–256.

https://doi.org/10.24355/dbbs.084-202102121510-0

[Martin, 2005] R. Martin, "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Aug. 2005.

[Martin and Breithaupt, 2003] R. Martin and C. Breithaupt, "Speech Enhancement in the DFT Domain Using Laplacian Speech Priors," in *Proc. of IWAENC*, Kyoto, Japan, Sep. 2003, pp. 87–90.

[McAulay and Malpass, 1980] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[Mowlaee and Saeidi, 2013] P. Mowlaee and R. Saeidi, "Target Speaker Separation in a Multisource Environment Using Speaker-Dependent Postfilter and Noise Estimation," in *Proc. of ICASSP*, Vancouver, BC, Canada, May 2013, pp. 7254–7258.

[Nahma et al., 2017] L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, "Convex Combination Framework for A Priori SNR Estimation in Speech Enhancement," in *Proc. of ICASSP*, New Orleans, LA, USA, Mar. 2017, pp. 4975–4979.

[NTT, 2012] "Super Wideband Stereo Speech Database," NTT Advanced Technology Corporation (NTT-AT), NTT Advanced Technology Corporation (NTT-AT), 2012.

[O'Shaughnessy, 1987] D. O'Shaughnessy, *Speech Communications: Human and Machine*. Reading, MA, USA: Addison-Wesley, 1987.

[Papamichalis, 1987] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Upper Saddle River, NJ, USA: Prentice Hall, Inc., 1987.

[Plapous et al., 2006] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[Rosenkranz, 2010] T. Rosenkranz, "Noise Codebook Adaptation for Codebook-Based Noise Reduction," in *Proc. of IWAENC*, Tel Aviv, Israel, Aug. 2010.

[Sigg et al., 2012] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech Enhancement Using Generative Dictionary Learning," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 20, no. 6, pp. 1698–1712, Aug. 2012.

[Simsekli et al., 2014] U. Simsekli, J. Le Roux, and J. R. Hershey, "Non-Negative Source-Filter Dynamical System for Speech Enhancement," in *Proc. of ICASSP*, Florence, Italy, May 2014, pp. 6206–6210.

[Srinivasan and Samuelsson, 2003] S. Srinivasan and J. Samuelsson, "Speech Enhancement

https://doi.org/10.24355/dbbs.084-202102121510-0

Using A-Priori Information," in *Proc. of Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1405–1408.

[Srinivasan et al., 2006] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook Driven Short-Term Predictor Parameter Estimation for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[Srinivasan et al., 2007] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[Stahl and Mowlaee, 2018] J. Stahl and P. Mowlaee, "A Simple and Effective Framework for A Priori SNR Estimation," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 5644–5648.

[Stylianou et al., 2007] Y. Stylianou, M. Faundez-Zanuy, and A. Eposito, *Progress in Nonlinear Speech Processing*. Berlin, Germany: Springer, 2007.

[Suhadi, 2012] S. Suhadi, "Speech Enhancement Using Data-Driven Concepts," Ph.D. dissertation, Technische Universität Braunschweig, 2012.

[Suhadi et al., 2011] S. Suhadi, C. Last, and T. Fingscheidt, "A Data-Driven Approach to A Priori SNR Estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[Taal et al., 2011] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[Talkin, 1995] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis*. New York: Elsevier Science Inc., 1995, pp. 495–518.

[Wolfe and Godsill, 2001] P. J. Wolfe and S. J. Godsill, "Simple Alternatives to the Ephraim and Malah Suppression Rule for Speech Enhancement," in *Proc. of SPWSSP*, Singapore, Singapore, Aug. 2001, pp. 496–499.

[Yegnanarayana and Murthy, 2000] B. Yegnanarayana and P. S. Murthy, "Enhancement of Reverberant Speech Using LP Residual Signal," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[Zhao and Kleijn, 2007] D. Y. Zhao and W. B. Kleijn, "HMM-Based Gain Modeling for Enhancement of Speech in Noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 882–892, Mar. 2007.

https://doi.org/10.24355/dbbs.084-202102121510-0

Publication I

S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "An Iterative Speech Model-Based A Priori SNR Estimator," in *Proc. of Interspeech*, Dresden, Germany, Sep. 2015, pp. 1740–1744

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

An Iterative Speech Model-Based A Priori SNR Estimator

Samy Elshamy[1]*, Nilesh Madhu*[2]*, Wouter Tirry*[2]*,*
Tim Fingscheidt[1]

[1]Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, D–38106 Braunschweig, Germany
[2]NXP Software
Interleuvenlaan 80, B–3001 Leuven, Belgium
{s.elshamy,t.fingscheidt}@tu-bs.de, {nilesh.madhu,wouter.tirry}@nxp.com

Abstract

In this contribution we propose an *a priori* signal-to-noise ratio (SNR) estimator based on a probabilistic speech model. Since the *a priori* SNR is an important means for speech enhancement algorithms, such as weighting rule calculation for noise reduction or speech presence probability computation, its diligent estimation is of wide interest. As a basis for this estimator a Gaussian mixture model (GMM) is trained on clean speech amplitudes and by finding the maximum likelihood (ML) clean speech estimate of the corresponding observed frame the *a priori* SNR can easily be calculated. Additionally, an iterative scheme is applied to consequently enhance the estimate by repetitively evaluating the GMM. This technique allows to accomplish noise reduction free of musical tones even in non-stationary noise environments and exceeds the quality of the classical decision-directed (DD) approach for typical spectral weighting rules.

Index Terms: speech enhancement, *a priori* SNR estimation

1. Introduction

The estimation of the *a priori* SNR has been subject to a number of publications in the past, e.g., [1, 2, 3, 4, 5]. It is an important entity for most speech enhancement applications such as spectral weighting rules for noise reduction algorithms [1, 2, 6, 7], means of speech presence probability estimation [8], and voice activity detection [9], for example. Since most of the weighting rules are functions of the *a priori* SNR and *a posteriori* SNR, with the *a priori* SNR having the stronger impact, it is important to have a good *a priori* SNR estimate at hand. The quality of the estimator influences the amount of introduced musical tones, speech distortion, and the degree of achieved noise suppression.

The most common and famous way is the DD approach by Ephraim and Malah [1] where a weighted sum of two components yields the desired estimate. Both components are representing *a priori* SNR-like entities where the first depends on previous estimates and the other on the current observation. Since the weights add up to one and the weight for the first component is mostly close to unity, the *a priori* SNR estimate is strongly influenced by the previous frame. As a consequence the DD approach deteriorates once sudden changes of the instantaneous *a priori* SNR occur [10, 11].

Martin et al. [3] propose a smoothing algorithm in the cepstral domain where the cepstrum of the ML *a priori* SNR is smoothed with an adaptive frequency bin-dependent smoothing factor which is adjusted depending on where the spectral envelope and the excitation is expected. Therefore the algorithm is relying on a fundamental frequency estimator. The authors are able to show that cepstral smoothing is superior to temporal smoothing since the cepstrum offers better abilities to individually smooth coefficients related to noise (stronger smoothing applied) and coefficients representing speech (weaker smoothing applied), and thus maintaining a better speech component while achieving high noise attenuation simultaneously.

Two data-driven approaches are introduced by Fingscheidt et al. [4] in 2011 where one utilizes two neural networks in order to estimate the *a priori* SNR on behalf of the two smoothing components of the DD approach. One network is trained under the hypothesis of speech presence and the other under speech absence. The outputs of the neural networks are finally smoothed with a smoothing factor based on an internal speech absence probability yielding the final *a priori* SNR estimate.

Our proposed approach is a continuation of the work of Mowlaee et al. [12] which can be interpreted as a noise reduction algorithm that we modify to provide an *a priori* SNR estimate. Our algorithm is designed as a two-step procedure where the first stage is a classical noise reduction composed of a noise power estimator, e.g., [13, 14, 15], an *a priori* SNR estimator, e.g., [1, 3], and finally one of the spectral weighting rules as already mentioned. Throughout the paper this preprocessing stage is denoted as preliminary noise reduction. As a second stage the ML clean speech amplitude estimate is selected by evaluating the preliminary enhanced signal with the appropriate metric against a GMM which has been trained on clean speech amplitudes. While Mowlaee et al. [12] directly deduce a simple spectral subtraction-related weighting rule from the ML clean speech estimate, we utilize the GMM for sole *a priori* SNR estimation and thus are independent from the employed weighting rule: We are able to use any *a priori* SNR-driven weighting rule to retrieve the clean speech estimate from the microphone signal. As an important improvement we introduce an iterative scheme where the GMM is repeatedly evaluated in order to retrieve a better ML clean speech amplitude for subsequent *a priori* SNR estimation.

The remainder of the publication is structured as follows: In Sec. 2 we introduce mathematical notations and briefly revisit the approach of Mowlaee et al. as we understand it. Along comes our generalized derivation of the *a priori* SNR by the GMM and the new iterative scheme. Next in Sec. 3 an evaluation of the *a priori* SNR estimator in combination with different spectral weighting rules is presented, and finally we draw our conclusions in Sec. 4.

https://doi.org/10.24355/dbbs.084-202102121510-0

2. Proposed *A Priori* SNR Estimation

Since our approach is derived from a paper describing a noise reduction we embed our contribution in a speech enhancement context as well. For completeness and better understanding we sketch the whole procedure of Mowlaee et al. in a slightly different manner. Note that the final spectral gain calculation and application can be omitted once we will use it for sole *a priori* SNR estimation.

2.1. Notations and Assumptions

We assume an additive noise model which can be expressed in the time domain by $y(n) = s(n) + d(n)$, with $y(n)$ being the microphone noisy speech signal, $s(n)$ the clean speech, and $d(n)$ the additive noise, while n denotes the discrete-time sample index. Applying the discrete Fourier transform (DFT) we obtain $Y_\ell(k) = S_\ell(k) + D_\ell(k)$, with frame index ℓ and frequency bin index k with $0 \leq k \leq K-1$. Additionally, we understand that the noise signal $d(n)$ can be split up into a stationary (ST) and a non-stationary (NST) component. Thus we define the noise signal as

$$\begin{aligned} d(n) &= d^{\text{NST}}(n) + d^{\text{ST}}(n) \\ y(n) - s(n) &= d^{\text{NST}}(n) + d^{\text{ST}}(n), \end{aligned} \tag{1}$$

leading to the frequency domain representation

$$Y_\ell(k) - S_\ell(k) = D_\ell(k) = D_\ell^{\text{NST}}(k) + D_\ell^{\text{ST}}(k). \tag{2}$$

Assuming statistical independence of speech and noise, as well as of the stationary and the non-stationary noise components, respectively, we obtain the power spectrum representation

$$|Y_\ell(k)|^2 - \sigma_S^2(\ell,k) = \sigma_D^2(\ell,k) = \sigma_D^{\text{NST}}(\ell,k)^2 + \sigma_D^{\text{ST}}(\ell,k)^2, \tag{3}$$

where the variance of the microphone signal has been replaced by its instantaneous estimate $\sigma_Y^2(\ell,k) = |Y_\ell(k)|^2$.

2.2. Model-Based Noise Reduction after Mowlaee

In the following we sketch the model-based speech enhancement approach of Mowlaee et al. [12] as we will partially adopt his algorithm.

A preliminary enhanced speech signal is obtained by employing a conventional noise reduction, here by means of the noise power estimator presented in [15], *a priori* SNR estimation by the DD approach from [1], and the minimum mean-square error log-spectral amplitude estimator (MMSE-LSA) [2]. Thus we can define the preliminary enhanced speech signal by

$$\bar{Y}_\ell(k) = Y_\ell(k) \cdot \bar{G}_\ell(k), \tag{4}$$

where $\bar{G}_\ell(k)$ represents the MMSE-LSA spectral gains [2], which are depending on the *a priori* SNR $\xi_\ell(k) = \frac{\sigma_S^2(\ell,k)}{\sigma_D^{\text{ST}}(\ell,k)^2}$, and the *a posteriori* SNR $\gamma_\ell(k) = \frac{|Y_\ell(k)|^2}{\sigma_D^{\text{ST}}(\ell,k)^2}$. Since some entities are not available as such, we need to rely on estimated quantities, whereby $\hat{\xi}_\ell(k)$ is provided by the DD *a priori* SNR estimator [1] and $\hat{\sigma}_D^{\text{ST}}(\ell,k)^2$ is the estimated noise power by [15]. Please note that the utilized noise power estimate is not perfectly capable of comprising the total noise power and thus a residual non-stationary noise component $\sigma_D^{\text{NST}}(\ell,k)^2$ still requires attention. Moreover the estimate of [15] is taken as $\hat{\sigma}_D^{\text{ST}}(\ell,k)^2$ even though it is supposed to cover non-stationary noise as well at least to some extent.

As a next step a binary mask is applied to the pre-enhanced signal in order to identify bins which contain active speech, yielding

$$\bar{Y}'_\ell(k) = \bar{Y}_\ell(k) \cdot B_\ell(k) \tag{5}$$

with

$$B_\ell(k) = \begin{cases} 1 & \text{, if } \sqrt{\hat{\sigma}_D^{\text{ST}}(\ell,k)^2} < |\bar{Y}_\ell(k)| \\ 0 & \text{, otherwise} \end{cases}. \tag{6}$$

Next we define a Gaussian mixture model (GMM) probability density function (PDF)

$$p(\mathbf{X}) = \sum_{m=1}^{M} c_m \cdot \mathcal{N}(\mathbf{X}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{7}$$

with M distinct modes, each representing a normal distribution $\mathcal{N}(\mathbf{X}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ with mean vector

$$\boldsymbol{\mu}_m = (\mu_m(0), \mu_m(1), \ldots, \mu_m(K-1))^T, \tag{8}$$

main-diagonal covariance matrix

$$\boldsymbol{\Sigma}_m = \text{diag}\left(\left(\sigma_m^2(0), \sigma_m^2(1), \ldots, \sigma_m^2(K-1)\right)\right), \tag{9}$$

and mixture weight c_m following the constraint $\sum_{m=1}^{M} c_m = 1$. Operator $(\cdot)^T$ denotes the transpose.

In a training step, the vector $\mathbf{X}$ represents clean speech DFT amplitudes according to $\mathbf{X} = (|S_\ell(0)|, |S_\ell(1)|, \ldots, |S_\ell(K-1)|)^T$. The GMM is trained by performing ten iterations with enforced main-diagonal covariance matrices of the expectation-maximization (EM) algorithm [16] on clean speech amplitudes.

The next step of the baseline approach under consideration is applying the pre-enhanced masked amplitudes (5)

$$\mathbf{X} = (|\bar{Y}'_\ell(0)|, |\bar{Y}'_\ell(1)|, \ldots, |\bar{Y}'_\ell(K-1)|)^T = \bar{\mathbf{X}}'_\ell \tag{10}$$

to each of the weighted modes in the GMM (7) and searching for the maximum likelihood (ML) clean speech amplitude estimate

$$\hat{\mathbf{X}}_\ell^{\text{ML}} = (|\hat{S}_\ell^{\text{ML}}(0)|, |\hat{S}_\ell^{\text{ML}}(1)|, \ldots, |\hat{S}_\ell^{\text{ML}}(K-1)|)^T = \boldsymbol{\mu}_{m^*}, \tag{11}$$

which is the mean vector of the Gaussian mode with index

$$\begin{aligned} m^* &= \arg\max_m c_m \cdot \mathcal{N}(\mathbf{X} = \bar{\mathbf{X}}'_\ell; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \qquad (12) \\ &= \arg\min_m \sum_{k=0}^{K-1} \left[\frac{\left(|\bar{Y}'_\ell(k)| - \mu_m(k)\right)^2}{2\sigma_m^2(k)} - \ln\left(\frac{c_m}{\sqrt{2\pi}\sigma_m(k)}\right) \right]. \end{aligned}$$

Now the standard deviation $\hat{\sigma}_D^{\text{NST}}(\ell,k)$ of the non-stationary noise power is estimated by a simple spectral power subtraction approach motivated by (3) and as presented in [12], where a Wiener filter-like solution is proposed to obtain the noise power estimate from the microphone signal $Y_\ell(k)$ by

$$\hat{\sigma}_D^{\text{NST}}(\ell,k) = |Y_\ell(k)| \cdot G_\ell^{\text{NST}}(k) \tag{13}$$

with

$$G_\ell^{\text{NST}}(k) = \frac{\left(|Y_\ell(k)|^2 - |\hat{S}_\ell^{\text{ML}}(k)|^2 - \hat{\sigma}_D^{\text{ST}}(\ell,k)^2\right)}{|Y_\ell(k)|^2} \tag{14}$$

which can be interpreted as (compare to (2)) $G_\ell^{\text{NST}}(k) = \frac{\hat{\sigma}_D^{\text{NST}}(\ell,k)^2}{\sigma_Y(\ell,k)^2}$. The final spectral gain is calculated on behalf of the maximum likelihood clean speech amplitude estimate (11), the stationary, and the non-stationary noise power estimates. The estimated frequency-domain clean speech signal $\hat{S}_\ell(k)$ is then

https://doi.org/10.24355/dbbs.084-202102121510-0

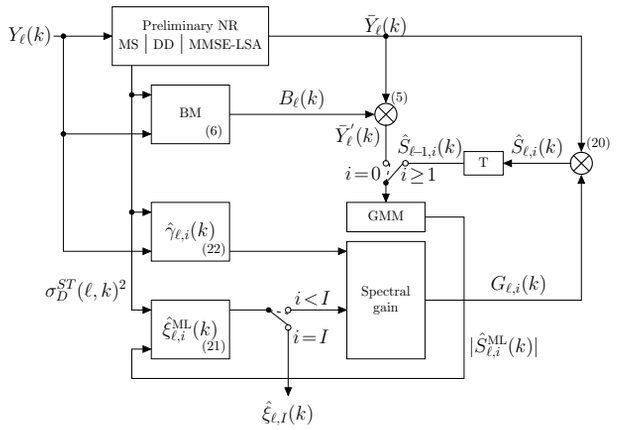


Figure 1: *Block diagram of the proposed iterative model-based a priori SNR estimator.*

obtained by

$$\hat{S}_\ell(k) = Y_\ell(k) \cdot G_\ell(k) \tag{15}$$

with $G_\ell(k) =$

$$\begin{cases} \frac{|\hat{S}_\ell^{\mathrm{ML}}(k)|}{\sqrt{|\hat{S}_\ell^{\mathrm{ML}}(k)|^2 + \max\left(\hat{\sigma}_D^{\mathrm{ST}}(\ell,k)^2, \hat{\sigma}_D^{\mathrm{NST}}(\ell,k)^2\right)}}, & \text{if } \sqrt{\hat{\sigma}_D^{\mathrm{ST}}(\ell,k)^2} < |\hat{S}_\ell^{\mathrm{ML}}(k)| \\ G_{\min} & \text{, otherwise} \end{cases}, \tag{16}$$

where $G_{\min}$ denotes a typical gain floor of -15 dB.

2.3. New Model-Based *A Priori* SNR Estimation

In (16) the model-based clean speech amplitude estimate is used to form a specific gain function. However, considering the existing algorithm pipeline, it is easy to see that one can influence a key factor: the *a priori* SNR. The *a priori* SNR, is not only important for noise reduction applications but also for speech processing in general. Instead of computing a particular weighting rule directly as done in (16) the sole *a priori* SNR widens the scope of application in speech enhancement, e.g., for speech presence detection. The block diagram of our proposed architecture is depicted in Fig. 1.

We introduce the new estimate for the *a priori* SNR based on the ML clean speech amplitude estimate as

$$\hat{\xi}_\ell^{\mathrm{ML}}(k) = \frac{|\hat{S}_\ell^{\mathrm{ML}}(k)|^2}{\hat{\sigma}_D^{\mathrm{ST}}(\ell,k)^2} \tag{17}$$

and also adopt the estimated *a posteriori* SNR

$$\hat{\gamma}_\ell(k) = \frac{|Y_\ell(k)|^2}{\hat{\sigma}_D^{\mathrm{ST}}(\ell,k)^2}. \tag{18}$$

Next, we consider the application of the gain function in (15) and rewrite (16) as

$$G_\ell(k) = \begin{cases} \sqrt{\frac{\hat{\xi}_\ell^{\mathrm{ML}}(k)}{1+\hat{\xi}_\ell^{\mathrm{ML}}(k)}}, & \text{if } \hat{\xi}_\ell^{\mathrm{ML}}(k) > 1 \\ G_{\min} & \text{, otherwise} \end{cases}. \tag{19}$$

Please note that in the derivation of (19) from (16) we disregard the max operation in the denominator of (16) for the moment and simply use $\hat{\sigma}_D^{\mathrm{ST}}(\ell,k)^2$ instead.

Having defined $\hat{\xi}_\ell^{\mathrm{ML}}(k)$ and $\hat{\gamma}_\ell(k)$ as in (17) and (18), respectively, we are now able to substitute the gain function $G_\ell(k)$ by any given weighting rule that depends on $\hat{\xi}_\ell^{\mathrm{ML}}(k)$ and/or $\hat{\gamma}_\ell(k)$, e.g., by the MMSE-LSA [2].

2.4. New Iterative Approach

Our iterative approach is based on the observation that the repetitive application of $G_{\ell,i}(k)$ to the preliminary enhanced signal $\bar{Y}_\ell(k)$ yields a more suitable signal w.r.t. *a priori* SNR estimation. This does not necessarily imply that the obtained intermediate clean speech estimates improve w.r.t. speech quality.

First, we define the iterative clean speech amplitude estimate based on (11) by introducing the iteration index i, yielding

$$\hat{\mathbf{X}}_{\ell,i}^{\mathrm{ML}} = (|\hat{S}_{\ell,i}^{\mathrm{ML}}(0)|, |\hat{S}_{\ell,i}^{\mathrm{ML}}(1)|, \ldots, |\hat{S}_{\ell,i}^{\mathrm{ML}}(K-1)|)^T = \boldsymbol{\mu}_{m_i^*} \tag{20}$$

with Gaussian mode index $m_i^* =$

$$\arg\min_m \sum_{k=0}^{K-1} \left[\frac{\left(|\hat{S}_{\ell,i-1}(k)| - \mu_m(k)\right)^2}{2\sigma_m^2(k)} - \ln\left(\frac{c_m}{\sqrt{2\pi}\sigma_m(k)}\right) \right] \tag{21}$$

for $i = 1, ..., I$, where I denotes the maximum number of performed iterations. We initialize $\hat{S}_{\ell,0}(k) = \bar{Y}'_\ell(k)$ and define an update rule for the estimated frequency-domain clean speech signal as follows

$$\hat{S}_{\ell,i}(k) = \bar{Y}_\ell(k) \cdot G_{\ell,i}(k) \tag{22}$$

with $G_{\ell,i}(k)$ being the MMSE-LSA spectral weighting rule [2] depending on the iterative *a priori* SNR

$$\hat{\xi}_{\ell,i}^{\mathrm{ML}}(k) = \frac{|\hat{S}_{\ell,i}^{\mathrm{ML}}(k)|^2}{\hat{\sigma}_D^{\mathrm{ST}}(\ell,k)^2} \tag{23}$$

and the fixed *a posteriori* SNR.

$$\hat{\gamma}_{\ell,i}(k) = \hat{\gamma}_\ell(k). \tag{24}$$

Since the iterative weighting rule is applied to the same signal $\bar{Y}_\ell(k)$ each time (22), we define a convergence criterion being that the same mode from the GMM is consecutively chosen: $m_i^* \stackrel{!}{=} m_{i-1}^*$ or a maximum amount of iterations reached. This implies that the gain function $G_{\ell,i}(k)$ does not change once the same amplitude mean has been selected by (21).

3. Experimental Results

3.1. Experimental Setup

Our framing throughout the whole training and testing process is the following: We operate at a sample rate of 8 kHz with a frame size of 256 samples and a frame shift of 64 samples. The employed window for analysis and synthesis is a periodic square root Hann window. Since the core of the approach is a GMM trained on clean speech amplitudes we need a significant amount of speech data. We chose the GRID corpus [17], downsampled to 8 kHz and divided it into a test set containing speakers 2, 4, 6, 7, and a training set comprising the remainder of the speakers. Thus we have two male and two female speakers in our test set and our test results are averaged over 100 sentences per speaker, i.e., 400 total. Since the files are alphabetically sorted and follow a certain grammar we designate every 10th file of every speaker to be in the test set in order to achieve some variance. The training of a speaker-independent GMM is based on the remaining 30 speakers and 50 files per speaker amounting to a total of 1500 files. Next we extract the clean

https://doi.org/10.24355/dbbs.084-202102121510-0

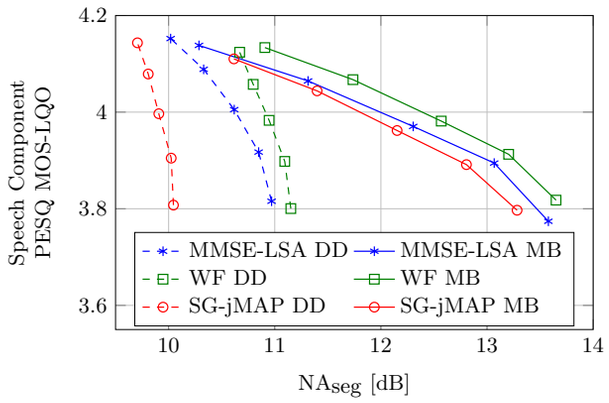


Figure 2: *Comparing the new a priori SNR estimator (MB) to the DD approach in combination with three different weighting rules by means of segmental NA and PESQ MOS of the speech component.*

speech amplitudes of the training pool by employing a simple 3-state voice activity detection as proposed in [18], please note that we do not perform any normalization on the source files. Subsequently the EM algorithm is employed with forced main-diagonal covariance matrices, a maximum of 100 iterations, and the desired order of $M = 512$. Hence the computed GMM contains 512 modes representing clean speech amplitudes, each being a vector with dimension corresponding to the DFT size $K = 256$.

For the simulation we make use of the ETSI background noise database [19], downsampled to 8 kHz, and employ the left channel noise files recorded in a full-size car driving at 80 km/h, in a call center, and at crossroads. Subsequently we measure the active speech level [20], adjust the level of the noise accordingly to obtain the desired SNR ranging from -5 dB to 15 dB in 5 dB steps, and superimpose speech and noise.

Evaluation is performed by employing a standard noise reduction using minimum statistics (MS) [14] as noise power estimator adjusted with a fixed overestimation factor for each weighting rule such that a comparable PESQ score is met at -5 dB SNR. We apply the new and the DD *a priori* SNR estimator, and three different weighting rules, namely MMSE-LSA [2], the Wiener filter (WF) [7], and Lotter's super-Gaussian joint maximum *a posteriori* (SG-jMAP) estimator [6]. The reference algorithms are operated with the optimal parameters[1] as proposed in [21]. Our setup of the preliminary noise reduction of the iterative *a priori* SNR estimator as described in Sec. 2.2 is as follows: The noise power is estimated by employing the (MS) algorithm, the smoothing factor for the DD *a priori* SNR estimation is set to $\beta_{\text{DD}} = 0.985$, and the minimum for the *a priori* SNR is $\xi_{\text{min}} = -15$ dB as is the gain floor G_{min} for the MMSE-LSA weighting rule.

As we assume a linear model in Sec. 2.1 the components of the noisy signal can be separately processed by applying the respective gain function $G_\ell(k)$ to the individual signals $s(n)$ and $d(n)$ in the frequency domain. This so-called white-box approach [22] yields the filtered clean speech $\tilde{s}(n)$ and the filtered noise $\tilde{d}(n)$, respectively.

For measures of quality we are taking into account the segmental noise attenuation (NA) computed as [23]

$$\text{NA}_{\text{seg}} = 10\log_{10}\left[\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\text{NA}(\ell)\right], \tag{25}$$

with

$$\text{NA}(\ell) = \frac{\sum_{\nu=0}^{N-1} d^2(\nu+\ell N)}{\sum_{\nu=0}^{N-1}\tilde{d}^2(\nu+\ell N+\delta)},$$

where ℓ defines a segment of length $N = 256$, δ is compensating the sample delay of the filtered signals, and $\frac{1}{|\mathcal{L}|}$ is a normalization factor since $|\mathcal{L}|$ represents the cardinality of set $\mathcal{L}$, containing all frames. Besides, we employ the PESQ MOS-LQO score [24] to evaluate the quality of the *filtered clean speech component* $\tilde{s}(n)$, *not* of the total enhanced speech $\hat{s}(n)$.

3.2. Experimental Evaluation

Fig. 2 shows the results of our simulations averaged for the three different noise environments. We plot the PESQ MOS scores over the segmental NA. Every marker denotes a different SNR condition, starting with 15 dB at the top to -5 dB at the bottom in steps of 5 dB. The dashed lines represent the reference *a priori* SNR estimator (DD) while the solid ones depict the proposed approach. Each *a priori* SNR estimator is evaluated in combination with the different weighting rules, MMSE-LSA, SG-jMAP, and WF, which are distinguished by the different markers. In general, the further to the top and to the right a curve is located in the plot, the lower the speech distortion and the amount of perceived residual noise in the enhanced speech signal. As we tuned all the different setups to achieve a similar PESQ score at -5 dB to facilitate a fair comparison, the curves only differ in the achieved level of NA while the quality of the speech component remains similar in terms of PESQ scores.

The figure shows that the proposed *a priori* SNR estimator allows to achieve a much higher segmental NA while maintaining comparable speech distortion (speech component PESQ scores). Especially in low SNR conditions the new approach outperforms the DD estimation by up to more than 4 dB segmental NA. Interestingly the model-based estimation leads to a far stronger decrease of segmental NA in higher SNR conditions but still exceeds the reference algorithms in all three cases. It is also observable that the performance of the different weighting rules seems to be less varying with the new approach since the curves show very similar behavior and exhibit less variance in the results when compared to the reference approach. Informal listening tests have also shown that when musical tones are present in an enhanced file processed with the DD estimator they vanish when alternative *a priori* SNR estimation is applied independent of the utilized weighting rule.

4. Conclusion

In this contribution we presented a new *a priori* SNR estimator based on a GMM providing a ML clean speech amplitude estimate. We have evaluated our approach against Ephraim/Mahla's DD estimator as reference, three different weighting rules, and three different noise environments on the GRID corpus. Our experiments have shown that the proposed estimator outperforms the DD approach, not only, but especially in low SNR conditions in terms of segmental NA while maintaining a comparable quality of the speech component.

[1] Parameters for the different weighting rules as applied for the evaluation:

MMSE-LSA: $\beta_{\text{DD}} = 0.975, \xi_{\text{min}} = -15$ dB
WF: $\beta_{\text{DD}} = 0.990, \xi_{\text{min}} = -14$ dB
SG-jMAP: $\beta_{\text{DD}} = 0.993, \xi_{\text{min}} = -14$ dB.

https://doi.org/10.24355/dbbs.084-202102121510-0

5. References

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[2] ——, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[3] C. Breithaupt, T. Gerkmann, and R. Martin, "A Novel A Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Las Vegas, NV, USA: IEEE, Mar. 2008, pp. 4897–4900.

[4] S. Suhadi, C. Last, and T. Fingscheidt, "A Data-Driven Approach to A Priori SNR Estimation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 1, pp. 186–195, Jan. 2011. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2045799

[5] I. Cohen, "Speech Enhancement Using Super-Gaussian Speech Models and Noncausal A Priori SNR Estimation," *Speech Communication*, vol. 47, no. 3, pp. 336–350, Nov. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2005.02.011

[6] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[7] P. Scalart and J.V. Filho, "Speech Enhancement Based on A Priori Signal To Noise Estimation," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2. Atlanta, GA, USA: IEEE, May 1996, pp. 629–632.

[8] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio with Fixed Priors," *IEEE Transactions on Audio Speech and Language Processing*, vol. 16, no. 5, pp. 910–919, July 2008.

[9] J. Sohn, N.S. Kim, and W. Sung, "A Statistical and Model-Based Voice and Activity Detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[10] O. Capp, "Elimination of the Musical and Noise Phenomenon and with the Ephraim and Malah Noise and Suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[11] C. Breithaupt and R. Martin, "Analysis of the Decision-Directed SNR Estimator for Speech Enhancement With Respect to Low-SNR and Transient Conditions," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2047681

[12] P. Mowlaee and R. Saeidi, "Target Speaker Separation in a Multisource Environment Using Speaker-Dependent Postfilter and Noise Estimation," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 2013, pp. 7254–7258.

[13] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003. [Online]. Available: http://dx.doi.org/10.1109/TSA.2003.811544

[14] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[15] S. Rangachari and P. C. Loizou, "A Noise-Estimation Algorithm for Highly Non-Stationary Environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, Feb. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2005.08.005

[16] G. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley & Sons, Inc., 2000.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, p. 2421, 2006. [Online]. Available: http://dx.doi.org/10.1121/1.2229005

[18] B. Fodor and T. Fingscheidt, "Reference-free SNR Measurement for Narrowband and Wideband Speech Signals in Car Noise," in *Proc. of ITG Conf. on Speech Communication*, Braunschweig, Germany, Sept. 2012, pp. 1–4.

[19] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ), Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database."

[20] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.

[21] H. Yu, "Post-Filter Optimization for Multichannel Automotive Speech Enhancement," Ph.D. dissertation, Technische Universität Braunschweig, 2013.

[22] S. Gustafsson, R. Martin, and P. Vary, "Proceedings of Workshop on Quality Assessment in Speech, Audio and Image Communication," in *Proc. of Workshop on Quality Assessment in Speech, Audio, and Image Communication*, Darmstadt, Germany, Mar. 1996, pp. 36–40.

[23] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-Optimized Speech Enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008. [Online]. Available: http://dx.doi.org/10.1109/TASL.2008.920062

[24] ITU, *Rec. P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Feb. 2001.

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

Publication II

S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "Instantaneous A Priori SNR Estimation by Cepstral Excitation Manipulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

Instantaneous *A Priori* SNR Estimation by Cepstral Excitation Manipulation

Samy Elshamy, Nilesh Madhu, Wouter Tirry, and Tim Fingscheidt, *Senior Member, IEEE*

***Abstract*—As the *a priori* signal-to-noise ratio (SNR) contains crucial information about a signal's mixture of speech and noise, its estimation is subject to steady research. In this paper, we introduce a novel *a priori* SNR estimator based on synthesizing an idealized excitation signal in the cepstral domain. Our approach utilizes a source-filter decomposition in combination with a cepstral excitation manipulation in order to recreate an idealized excitation, which is subsequently shaped by an immanent envelope. In contrast to the well-known decision-directed approach by Ephraim and Malah, an *instantaneous* estimate is obtained, which is less prone to sudden acoustic environmental changes and musical noise. Additionally, the proposed estimator is able to preserve weak harmonic structures resulting in a spectrum that is more full-bodied. We present both a speaker-independent and a speaker-dependent variant of the new *a priori* SNR estimator, both showing more than 2 dB ΔSNR improvement versus state of the art, without any significant increase in speech distortion.**

***Index Terms*—*A priori* SNR, speech enhancement.**

I. Introduction

A priori SNR estimation has long been an important topic in speech enhancement. Having only a single mixture at hand most likely impedes enhancement tasks since no knowledge about the individual components of the observed mixture is available. Consequently, the need to estimate an *a priori* SNR arises and has been subject to research in several publications [1]–[7]. Algorithms such as voice activity detection [8], speech presence probability estimation [9] and, most importantly, spectral weighting rules for noise reduction algorithms [1], [2], [10], [11] take great profit from reliable *a priori* SNR estimates.

The decision-directed (DD) approach to estimate the *a priori* SNR by Ephraim and Malah [1] has been published along with a spectral amplitude estimator for noise reduction and is basically a weighted sum of two components. The first component is depicting the ratio of the previous frame's squared clean speech amplitude estimate and the provided noise power estimate also taken from the previous frame. The second component is an instantaneous estimate derived from the current frame's *a posteriori* SNR. The weights of both components sum up to unity and as proposed in [1] the weight for the first component is chosen close to unity. The approach has been thoroughly analyzed in [12] and [13], where the analysis of Cappé [12] has shown, that the DD *a priori* SNR estimate follows the *a posteriori* SNR[1] with a delay of one frame.

Cohen proposed a non-causal estimator which buffers a few frames and thus is capable of differentiating between onsets of speech and bursts of noise allowing less musical tones and distortion of transient speech regions [3]. The approach is also less sensitive to changes in the underlying speech model compared to the DD technique. In practice all this comes at the price of some frames of delay.

Breithaupt *et al.* proposed an estimator [5] which employs a quefrency-selective smoothing of a maximum likelihood (ML) speech power spectral density derived from the *a posteriori* SNR and the noise power estimate. In the cepstral domain the coefficients corresponding to the excitation and the envelope are smoothed differently. As a result they obtain an *a priori* SNR estimate that yields better results in a noise suppression framework than the one in [1] w.r.t. spectral distortion and musical tones, especially in non-stationary environments. However, the clean excitation is not directly modeled, still leaving potential for improvement.

A data-driven approach based on the DD formula has been published in [6]. The two components of the weighted sum are both input to two different neural networks, discriminating speech active and inactive frames, with the ideal *a priori* SNR as a target during the training process. In a practical system both networks are evaluated and a linear combination of the provided outputs yields the final *a priori* SNR estimate. The authors are able to show a reduction of speech distortion during speech onsets while maintaining a high noise attenuation during speech pause. As the training process requires noise signals the approach is not entirely independent of the noise type.

Our latest work [7] shows that a simple Gaussian mixture model (GMM), representing clean speech spectral amplitudes, is able to provide a ML clean speech amplitude estimate when preliminary denoising is applied to the observation and subsequently the GMM is evaluated. The provided estimate is then used as numerator for an intermediate *a priori* SNR estima-

Manuscript received October 20, 2016; revised March 15, 2017 and April 28, 2017; accepted April 30, 2017. Date of publication May 9, 2017; date of current version June 12, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mads Græsbøll Christensen. *(Corresponding author: Tim Fingscheidt.)*

S. Elshamy and T. Fingscheidt are with the Institute for Communications Technology, Technische Universität Braunschweig, 38106 Braunschweig, Germany (e-mail: s.elshamy@tu-bs.de; t.fingscheidt@tu-bs.de).

N. Madhu and W. Tirry are with the NXP Software, 3001 Leuven, Belgium (e-mail: nilesh.madhu@nxp.com; wouter.tirry@nxp.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2702385

[1]Please note that the term "*a posteriori* SNR" in [12] differs from its use in mainstream literature as introduced in [1].

https://doi.org/10.24355/dbbs.084-202102121510-0

tion and continuously improved by repeated filtering and re-estimation.

Motivated by Cappé's observation, Plapous *et al.* propose a so-called two-step noise reduction (TSNR) technique [14] which is able to compensate for the one-frame delay. It is used as a preliminary noise reduction for their harmonic regeneration noise reduction (HRNR) introduced in [4]. The HRNR approach employs an improved *a priori* SNR estimator which applies a non-linear function to an enhanced time-domain signal in order to restore lost harmonics in the spectrum. The enhanced signal is subsequently mixed with the preliminary denoised signal, according to the calculated gains of the TSNR, and then used as numerator for the *a priori* SNR estimate. The applied non-linearity produces an unnatural harmonic and leads to audible artifacts in certain low-frequency noise types.

A recent analysis [15] deals with the over- and underestimation of estimated *a priori* SNR. The authors propose to use a correction term based on an empirically obtained distribution of the true bias in dependency of the *a priori* and *a posteriori* SNRs. The distribution is then subject to a vector quantizer which is later on used to estimate the bias on real data to compensate for the aberration. They show how to improve the DD and also the TSNR approach and additionally state that the proposed method could be used together with any spectral weighting rule.

In this paper we introduce a novel approach that consequently exploits the a priori knowledge that comes along with a model-based approach, while staying fully independent of noise types. The proposed method for *instantaneous a priori* SNR estimation without the need for lookahead is based on the *source-filter* model representing human speech production and also embraces the convenience a cepstral representation offers in terms of pitch estimation and cosine synthesis.

Furthermore, we also address a problem known to occur with approaches that model solely the spectral shape as they typically lack the fine structure of the speech spectrum and thus are not able to suppress noise between the harmonics [16].

In a first stage we employ a preliminary noise reduction driven by a noise power estimator such as [17]–[19], suitable state-of-the-art *a priori* SNR estimation [1], [5]–[7], and a weighting rule of choice, e.g., [1], [2], [10], [11]. In a second stage we utilize linear predictive coding (LPC) analysis to decompose the preliminary denoised signal into its spectral envelope and excitation followed by a transformation of the excitation signal to the cepstral domain. Subsequently, we detect the pitch, and, as a core of our approach, we synthesize an *idealized excitation* which is shaped by the spectral envelope of the preliminary denoised signal. The resulting spectrum is finally used as clean speech amplitude estimate for an instantaneous *a priori* SNR numerator. We then extend our approach to using two variants of excitation manipulation (synthetic and template-based) and show improvement of the template-based over the purely synthetically created excitation. Finally, we investigate the potential of a speaker-dependent (vs. a speaker-independent) setup of our estimator.

The structure of this paper is as follows: In Section II we introduce our mathematical notations and some baseline estimators, which serve as reference in the evaluation. Next, we present our cepstral processing methodology in Section III followed by the two proposed manipulation schemes in Section IV. In Section V we present the experimental results and discussion separately, and conclude the paper in Section VI.

II. Notations and Baselines

We assume an additive model for the microphone signal $y(n)$ in the time domain as

$$y(n) = s(n) + d(n), \tag{1}$$

with $s(n)$ being the clean speech signal we are interested in, and $d(n)$ being the noise signal we aim to suppress. The discrete-time sample index is n. The corresponding frequency-domain representation by applying the discrete Fourier transform (DFT) is

$$Y_\ell(k) = S_\ell(k) + D_\ell(k), \tag{2}$$

with frame index ℓ and frequency bin index k being restricted by the DFT size K to $0 \leq k \leq K-1$. Furthermore, as most approaches do, we assume that the speech and noise signals are zero-mean and statistically independent of one another.

A. Noise Power Estimation

An estimate of the noise power, which is denoted by $\hat{\sigma}_\ell^D(k)^2$, is required for noise reduction and can be obtained by several algorithms which have been published in the past. Among those is the minimum statistics (MS) approach [17], which is a commonly utilized estimator with good performance in stationary and non-stationary environments. Besides, there are further estimators such as the minima-controlled estimator proposed in [18], or estimators based on the minimum mean-square error (MMSE), e.g., [20].

B. Spectral Weighting Rules

The desired clean speech spectral estimate is generally obtained by applying a real-valued gain function, also referred to as spectral weighting rule $G_\ell(k)$, to the observed signal as follows

$$\hat{S}_\ell(k) = Y_\ell(k) \cdot G_\ell(k). \tag{3}$$

Thereby, the noisy phase is usually maintained as motivated in [1], [21], [22], although some more recent publications support phase-aware speech enhancement [23]–[25]. As the potential of amplitude-based speech enhancement seems not yet exhausted, we feel comfortable to focus on these in the following.

Amongst the most famous weighting rules utilized to calculate gain functions $G_\ell(k)$, we find the well-known Wiener filter (WF) [10], the MMSE short-time spectral amplitude estimator (MMSE-STSA) [1], the MMSE log-spectral amplitude estimator (MMSE-LSA) [2], and the super-Gaussian joint maximum a posteriori (SG-jMAP) estimator [11]. The aforementioned various frequency bin-selective gain functions $G_\ell(k)$ are mostly (nonlinear) functions $f(\cdot)$ of the *a priori* SNR

$$\xi_\ell(k) = \frac{\sigma_\ell^S(k)^2}{\sigma_\ell^D(k)^2} \tag{4}$$

https://doi.org/10.24355/dbbs.084-202102121510-0

and partly also of the *a posteriori* SNR

$$\gamma_\ell(k) = \frac{|Y_\ell(k)|^2}{\sigma_\ell^D(k)^2} \tag{5}$$

allowing us to compute $G_\ell(k)$ as:

$$G_\ell(k) = f\left(\xi_\ell(k), \gamma_\ell(k)\right). \tag{6}$$

Since both entities require quantities that are not available in practice they need to be estimated (or at least components of them). We denote estimated entities with a hat $\hat{\cdot}$ as accent.

C. A Priori SNR Estimation

In this section we briefly sketch three baseline approaches which will later serve to compare our approach against.

1) Decision-Directed (DD)*:* The historic breakthrough to estimate the *a priori* SNR is the already mentioned DD approach by Ephraim and Malah [1]. In summary, the DD formula narrows down to

$$\hat{\xi}_\ell^{\mathrm{DD}}(k) = \\ (1-\beta_{\mathrm{DD}}) \cdot \max\left\{\hat{\gamma}_\ell(k) - 1, 0\right\} + \beta_{\mathrm{DD}} \frac{|\hat{S}_{\ell-1}(k)|^2}{\hat{\sigma}_{\ell-1}^D(k)^2}, \tag{7}$$

with β_{DD} and $(1-\beta_{\mathrm{DD}})$ being the weights of both components as mentioned in the introduction. Subsequently, as proposed in [12], the *a priori* SNR estimate is lower-bounded to a certain $\xi_{\min}$ to avoid musical tones.

2) Selective Cepstro-Temporal Smoothing (CTS)*:* This method, proposed by Breithaupt *et al.* [5], is utilizing properties of the cepstral representation to obtain a more precise *a priori* SNR estimate. The core of this approach is an adaptive, first-order recursive smoothing of the cepstrum of the ML clean speech estimate $c_\ell^{\hat{S}^{\mathrm{ML}}}(m)$ according to

$$\bar{c}_\ell^{\hat{S}}(m) = \alpha_\ell(m) \cdot \bar{c}_{\ell-1}^{\hat{S}}(m) + (1-\alpha_\ell(m)) \cdot c_\ell^{\hat{S}^{\mathrm{ML}}}(m), \tag{8}$$

where $\bar{c}_\ell^{\hat{S}}(m)$ is the smoothed version of the cepstrum and $m \in \mathcal{M} = \{0, 1, \ldots, K-1\}$ is the cepstral bin index. The cepstrum of the ML clean speech estimate in this particular case is obtained as

$$\left(c_\ell^{\hat{S}^{\mathrm{ML}}}(m)\right)_{m=0}^{K-1} = \mathrm{IDFT}\left\{\left(\log|\hat{S}_\ell^{\mathrm{ML}}(k)|^2\right)_{k=0}^{K-1}\right\}, \tag{9}$$

with

$$|\hat{S}_\ell^{\mathrm{ML}}(k)|^2 = \hat{\sigma}_\ell^D(k)^2 \cdot \max\left\{\xi_\ell^{\mathrm{ML}}(k), \xi_{\min}^{\mathrm{ML}}\right\}. \tag{10}$$

The ML *a priori* SNR floor $\xi_{\min}^{\mathrm{ML}} > 0$ is a small number yielding numerical stability, while

$$\xi_\ell^{\mathrm{ML}}(k) = \gamma_\ell(k) - 1, \tag{11}$$

as shown in [5]. Parameter $\alpha_\ell(m)$ is not only time-variant, but also quefrency-selective. Cepstral coefficients with small indices controlling the shape of the spectral envelope are to be smoothed only slightly, whereas the higher-indexed cepstral coefficients are supposedly related to noise and thus heavily smoothed. An exception is made for bins related to the fundamental frequency as these are suggested to be smoothed even less than the envelope-related quefrencies. Therefore, this method relies on a cepstral pitch estimation. For the detailed smoothing scheme we refer to [5]. After a bias compensation required due to the smoothing in the logarithmic domain and inverse transformation the final *a priori* SNR estimate is obtained as

$$\hat{\xi}_\ell^{\mathrm{CTS}}(k) = \max\left\{\frac{|\hat{S}_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}, \xi_{\min}\right\}, \tag{12}$$

with

$$\left(|\hat{S}_\ell(k)|^2\right)_{k=0}^{K-1} = \exp\left(\kappa + \mathrm{DFT}\left\{\left(\bar{c}_\ell^{\hat{S}}(m)\right)_{m=0}^{K-1}\right\}\right), \tag{13}$$

and κ being a log-amplitude spectrum bias compensation. In our simulations an improved bias compensation term has been used as presented in [26]. The noise power estimation is not further restricted to any specific method. Finally, note that an instantaneous extension to CTS could be employed as being done in [27].

3) Harmonic Regeneration (HRNR)*:* The HRNR approach by Plapous *et al.* [4] is based on the DD estimator, but taking Cappé's observation into account to compensate for the one-frame delay of the *a priori* SNR, underlying some preliminary spectral weights $G_\ell^{\mathrm{DD}}(k)$. The authors employ a two-step noise reduction technique (TSNR) to accomplish the delay compensation. Therefore, they introduce a second gain function

$$G_\ell^{\mathrm{TSNR}}(k) = f(\hat{\xi}_\ell^{\mathrm{TSNR}}(k), \hat{\gamma}_\ell(k)) \tag{14}$$

with an updated *a priori* SNR

$$\hat{\xi}_\ell^{\mathrm{TSNR}}(k) = \frac{|Y_\ell(k) \cdot G_\ell^{\mathrm{DD}}(k)|^2}{\hat{\sigma}_\ell^D(k)^2} \tag{15}$$

being responsible for the actual compensation. A harmonic *spectral* regeneration method operates on the TSNR-enhanced signal $Y_\ell(k) \cdot G_\ell^{\mathrm{TSNR}}(k)$, applying a simple non-linear function in the time domain, here half-wave rectification, and thereby boosting the harmonics of voiced frames. After transformation the spectrum is depicted as $\check{S}_\ell(k)$, which is not directly used for clean speech estimation but for another *a priori* SNR estimate $\hat{\xi}_\ell^{\mathrm{HRNR}}(k)$. To obtain this estimate, $\check{S}_\ell(k)$ is mixed with the TSNR-enhanced signal according to the corresponding gain function as follows

$$\hat{\xi}_\ell^{\mathrm{HRNR}}(k) = \\ \frac{\alpha_\ell(k) \cdot |Y_\ell(k) \cdot G_\ell^{\mathrm{TSNR}}(k)|^2 + (1-\alpha_\ell(k)) \cdot |\check{S}_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2} \tag{16}$$

where the authors propose to use weights $\alpha_\ell(k) = G_\ell^{\mathrm{TSNR}}(k)$.

This constitutes the final *a priori* SNR estimate; again, the noise power estimator can be chosen from available literature for each of the proposed stages.

Throughout this paper we refer to a system that is composed of a noise power estimator (Section II-A), an *a priori* SNR estimation (Section II-C), and a spectral weighting rule (Section II-B) as either a common or a preliminary noise reduction.

https://doi.org/10.24355/dbbs.084-202102121510-0

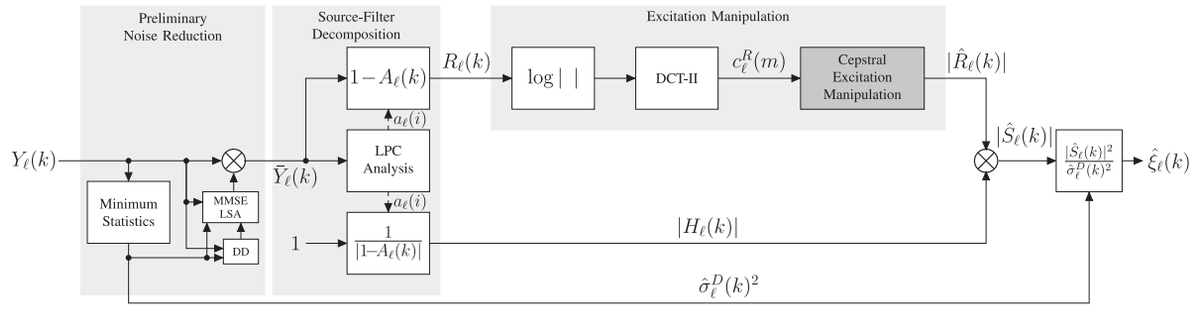


Fig. 1. Block diagram of the cepstral processing framework for our **proposed** *a priori* **SNR estimation**. The preliminary noise reduction consists of the MS noise power estimation algorithm, the DD *a priori* SNR estimation approach, and the MMSE-LSA spectral weighting rule.

III. New Cepstral Processing Framework for A Priori SNR Estimation

In this section we present the cepstral processing framework of our *a priori* SNR estimator, and provide some motivation for cepstral domain processing. Fig. 1 depicts its overall architecture.

A. Preliminary Noise Reduction

Similar to [4], we also employ a preliminary denoising stage before applying the actual approach. The motivation is to facilitate the extraction of required information for the proposed *a priori* SNR estimation. As it will rely on a pitch estimation, our approach benefits from the preliminary noise reduction rendering pitch estimation more robust, even in very low-SNR conditions. We target the preservation of harmonics which are often strongly attenuated, especially in adverse environments. In practice, this preliminary noise reduction is not limited to any specific components or approaches, but we propose to rely on a common noise reduction scheme being composed of some noise power estimation (e.g., minimum statistics [17]), the DD approach to *a priori* SNR estimation [1], and the MMSE-LSA spectral weighting rule [2] (referring to the left light gray block in Fig. 1).

B. Source-Filter Decomposition

Decomposing the preliminary denoised signal into its envelope and excitation (e.g., by LPC analysis) allows us to break down the enhancement task into two individual problems, thus enabling specific enhancement methods to be applied to each of the components, separately. However, in this paper we focus on the excitation only.

Our proposed method exploits knowledge about the process of human speech generation, especially of voiced speech. Therefore, it is important to have a reliable pitch estimation which is on the one hand supported by the preliminary denoising stage, and on the other hand by analyzing the excitation signal. For these reasons, we decompose the preliminary denoised signal $\bar{Y}_\ell(k)$ into the spectral excitation $R_\ell(k)$ and its spectral envelope $H_\ell(k)$, which is understood as the source and the filter, respectively. In each frame ℓ the spectral envelope $H_\ell(k)$ is obtained by first applying the K-point inverse discrete Fourier transform (IDFT) to the squared magnitude spectrum $|\bar{Y}_\ell(k)|^2$, resulting in the sequence of autocorrelation coefficients

$$\left(\varphi_\ell^{\bar{y},\bar{y}}(\nu)\right)_{\nu=0}^{K-1} = \text{IDFT}\left\{\left(|\bar{Y}_\ell(k)|^2\right)_{k=0}^{K-1}\right\}. \tag{17}$$

The first $N+1 < K$ elements $\varphi_\ell^{\bar{y},\bar{y}}(\nu)$, $\nu \in \{0,1,\ldots,N\}$ are used to compute a set of N LPC coefficients $a_\ell(i)$, $i \in \{1,2,\ldots,N\}$ by the Levinson-Durbin recursion. The LP analysis filter in the DFT domain $(1 - A_\ell(k))$ is then simply obtained by applying the K-point DFT to a sequence of the previously calculated N LPC coefficients, padded with $K - N - 1$ zeros:

$$\left(A_\ell(k)\right)_{k=0}^{K-1} = \text{DFT}\left\{(0, a_\ell(1), \ldots, a_\ell(N), 0, \ldots, 0)\right\}. \tag{18}$$

The LP analysis filter is employed to process the preliminary denoised signal $\bar{Y}_\ell(k)$ to retrieve the respective residual signal as [28]:

$$R_\ell(k) = \bar{Y}_\ell(k) \cdot \left(1 - A_\ell(k)\right), \tag{19}$$

while the spectral envelope is given by the inverse filter as

$$H_\ell(k) = \frac{1}{1 - A_\ell(k)}. \tag{20}$$

LPC analysis is an established method for source-filter decomposition [28]. An alternative to computing the envelope could have been simple liftering in the cepstral domain (i.e., taking only the lower part of the cepstrum), which, however, does not provide the exact same results as the Levinson-Durbin recursion of LPC analysis [29, Sec. 9.5.1].

Further investigations towards the processing framework as shown in Fig. 1 have also shown that a residual signal obtained via LPC analysis, being subsequently transformed into the cepstral domain, is better suited for later manipulation. This is further elaborated on in Section IV-B.

C. Cepstral Excitation Representation

Next, we obtain a cepstral representation of a signal by applying the discrete cosine transform of type II (DCT-II), but also an IFFT could have been chosen as in [5]. Additionally, we present some of its inherent, convenient properties we take advantage of. To further analyze the spectrum of the excitation signal in a first step we compute the cepstral coefficients upon the excitation signal's logarithmic magnitude spectrum as [30]

$$c_\ell^R(m) = \sum_{k=0}^{K-1} \log\left(|R_\ell(k)|\right) \cdot \cos\left[\pi m \left(k + 0.5\right) \tfrac{1}{K}\right] \tag{21}$$

https://doi.org/10.24355/dbbs.084-202102121510-0

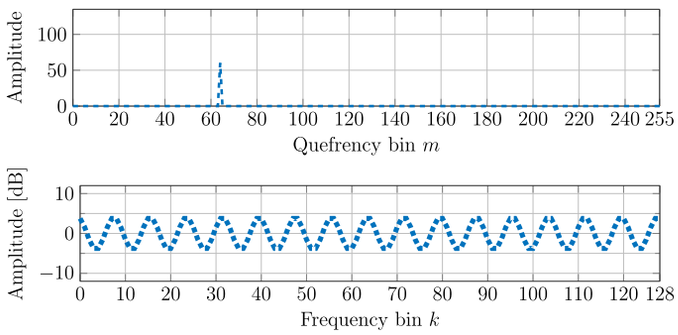


Fig. 2. Example of an **idealized synthetic excitation** for $K = 256$. Upper panel: Cepstrum $c_\ell^R(m)$ with $m_{F_0} = 64$ and $c_\ell^R(64) = 60$, representing a single zero-mean cosine in the log-spectral domain. Lower panel: Log-spectrum $20\log_{10}|R_\ell(k)|$ according to (22), showing only $\frac{K}{2}+1$ bins.

with $m \in \mathcal{M} = \{0, 1, \ldots, K-1\}$. The obtained cepstrum has a doubled resolution since we compute it on the whole spectrum (and not only on $\frac{K}{2}+1$ bins). The inverse DCT-II (IDCT-II) will be required in a later stage at the end of the cepstral excitation manipulation (CEM) and is calculated as

$$|\hat{R}_\ell(k)| = \exp\left(\frac{c_\ell^{\hat{R}}(0)}{K} + \frac{2}{K}\sum_{m=1}^{K-1} c_\ell^{\hat{R}}(m)\cdot\cos\left[\pi m\,(k+0.5)\,\tfrac{1}{K}\right]\right). \tag{22}$$

After the manipulations, the residual signal is mixed with the spectral envelope of the preliminary denoised signal as

$$|\hat{S}_\ell(k)| = |\hat{R}_\ell(k)|\cdot|H_\ell(k)| \tag{23}$$

which is used as numerator for the final *a priori* SNR estimate in an *instantaneous* fashion as follows

$$\hat{\xi}_\ell(k) = \frac{|\hat{S}_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}. \tag{24}$$

One of the most important properties of the cepstrum is the possibility to find a quefrency corresponding to the pitch by simple peak picking [31]. Thus, we estimate the pitch bin index m_{F_0} in a very naïve way by a maximum search of the cepstrum in a defined range specified by naturally occurring pitch values. Our focus is restricted to pitch frequencies F_0 from about 50 Hz to 500 Hz. Using[2] $f = \frac{2f_s}{m}$, the resulting cepstral bin indices at sampling frequency $f_s = 8$ kHz are therefore in the range $m \in \mathcal{M}_{F_0} = \{m_{500} = 32, \ldots, m_{50} = 320\}$. Pitch estimation on the basis of the residual signal after preliminary noise reduction is then simply performed according to

$$m_{F_0} = \arg\max_{\mu\in\mathcal{M}_{F_0}} \left(c_\ell^R(\mu)\right). \tag{25}$$

A further convenience is now the ability to easily synthesize a cosine in the log-spectral domain, by creating a cepstrum with only one non-zero bin. An example of such an *idealized* synthetic excitation is given in Fig. 2. The idea behind it is the fact that in voiced speech production harmonics occur at multiples of the fundamental frequency F_0, starting at F_0. A cosine in the log-spectral domain models this quite well already, as the maxima are located directly at the fundamental frequency and due to the periodicity also at the harmonics.

[2]The factor 2 stems from the doubled resolution of our cepstrum definition (21).

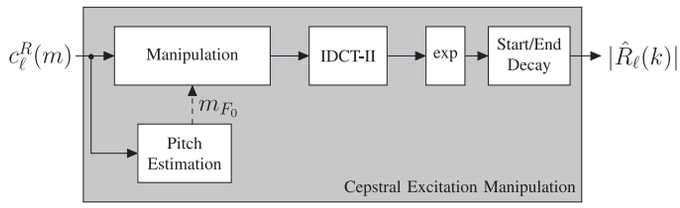


Fig. 3. Block diagram of the **proposed cepstral excitation manipulation** based on an **idealized synthetic excitation**.

IV. Cepstral Excitation Manipulation (CEM)

In the following, we introduce ways to manipulate the excitation in the cepstral domain (referring to the upper right gray block in Fig. 1). These methods form the core of our proposed approach. First, a manipulation towards an idealized *synthetic* excitation is introduced and second, a *template-based* alternative is presented.

A. *Idealized Synthetic Excitation* ($\mathrm{CEM_{ID}}$)

The first option we propose to manipulate the excitation in the cepstral domain is to completely replace it by an idealized synthetic one, followed by the IDCT-II (22) and some final manipulation of the start and the end of the log-spectrum (see Fig. 3). Having found the index m_{F_0} of the cepstral peak amplitude which corresponds to the pitch according to (25), we overestimate its amplitude and transfer it into our synthetic cepstrum

$$c_\ell^{\hat{R}}(m_{F_0}) = c_\ell^R(m_{F_0})\cdot\alpha_\ell(m_{F_0}), \tag{26}$$

while the remaining quefrencies, except for $(m = 0)$, are assigned a zero amplitude:

$$c_\ell^{\hat{R}}(m) = 0, \qquad \forall m \notin \{0, m_{F_0}\}. \tag{27}$$

In order to retain the energy of the preliminary denoised signal's residual, we preserve the cepstral energy coefficient $(m = 0)$:

$$c_\ell^{\hat{R}}(0) = c_\ell^R(0). \tag{28}$$

The proposed overestimation factor $\alpha_\ell(m) \geq 1$ could be time-variant and cepstral bin-dependent. While Fig. 2, upper panel, shows an example cepstrum, Fig. 4, upper panel, depicts the same cepstrum with applied cepstral overestimation factor. The resulting effect on the log-spectrum can be seen in Fig. 4, center panel, where a one-view comparison of both log-spectra is provided. Now, the benefit of the directed amplitude overestimation becomes obvious: The overestimation allows a narrower modeling of the harmonics (positive half waves), and also a correspondingly strong emphasis of the valleys (negative half waves) resulting in an increased attenuation between the harmonics. Note that this effect would not be obtained when boosting the harmonics in a shaped or already power-adjusted spectrum with a simple overestimation factor, as this would result only in a shift of the spectrum leaving the negative half

https://doi.org/10.24355/dbbs.084-202102121510-0

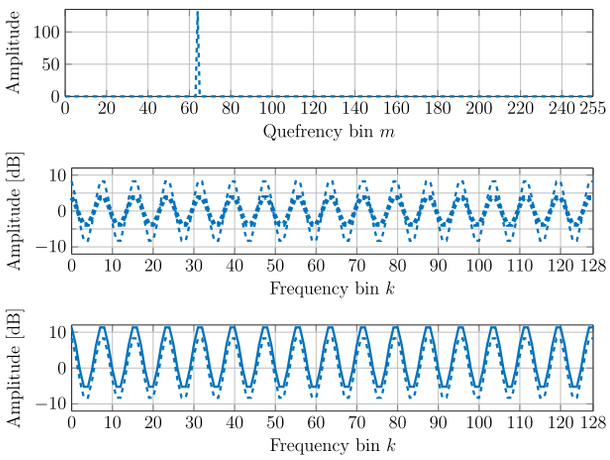


Fig. 4. Example of an **idealized synthetic excitation with overestimation** factor $\alpha_\ell(m_{F_0}) = 2.2$, DFT length $K = 256$, and optionally with preserved cepstral energy coefficient. Upper Panel: Cepstrum $c_\ell^{\hat{R}}(m)$ with $m_{F_0} = 64$ and $c_\ell^{\hat{R}}(64) = 60 \cdot 2.2 = 132$, representing a single zero-mean cosine in the log-spectral domain. Center Panel: Log-spectra $20\log_{10}|R_\ell(k)|$ (bold, dotted line from Fig. 2, lower panel) and $20\log_{10}|\hat{R}_\ell(k)|$ (dashed line, using (26)). Lower panel: Log-spectra $20\log_{10}|\hat{R}_\ell(k)|$ from center panel (dashed line), and power-adjusted log-spectra with additional $c_\ell^{\hat{R}}(0) = 90$ (solid line). All log-spectra show only $\frac{K}{2}+1$ bins.

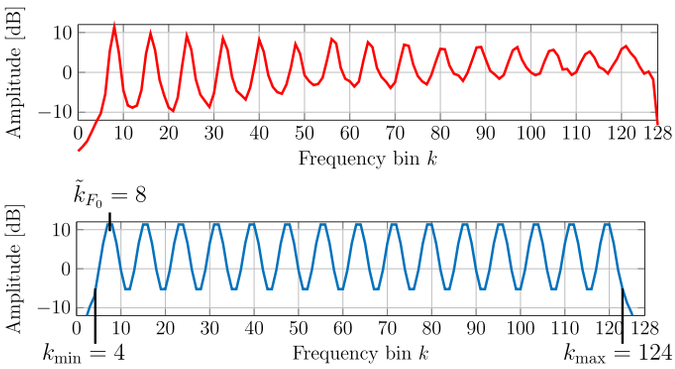


Fig. 5. Upper panel: Example of a log-spectrum excitation template for $m_{F_0} = 64$, obtained as described in Section IV-B. Lower panel: Example of an idealized synthetic excitation log-spectrum with preserved cepstral energy coefficient, overestimation factor $\alpha_\ell(m_{F_0}) = 1.5$, and applied start and end decay to remove the two false half periods.

waves unmodified. Besides, since our manipulation is in the cepstral domain, our approach translates consistently to all harmonics. This effect is difficult to achieve when operating in the spectral domain.

An overestimation of the energy coefficient would result in a scaling of the whole spectrum which is not desired here, as explained above. An example spectrum depicting the effect of (28) is shown in Fig. 4, lower panel, solid plot.

Naturally, spectral content of voiced human speech starts to occur at the fundamental frequency (after one period of the cosine), then being followed by the corresponding harmonics at multiples of the fundamental frequency, but there should be no spectral content prior to the fundamental frequency. Motivated by our observations during the training of excitation templates, we assume a similar effect at high frequencies (see Fig. 5, upper panel). Thus, a continuation of the cosine beyond the highest, fully representable harmonic is also not desired.

Similarly, in the HRNR approach [4, Fig. 8] a continuous harmonic log-amplitude spectrum is obtained. A problem that has been left unattended there is the falsely introduced half period at low frequencies, which is caused by the non-linear function in [4], applied in order to regenerate harmonics.

To tackle this issue, we propose a simple continuation of the decay of the cosine at low and high frequencies (instead of Fig. 4, lower panel, solid line, now Fig. 5, lower panel). To identify the first local minimum prior to the fundamental frequency we utilize $f = \frac{2f_s}{m}$ to obtain the corresponding pitch frequency F_0 based on the estimated cepstral bin index m_{F_0} from (25). We now convert the pitch frequency F_0 to its corresponding real-valued frequency-domain "bin index" as $\tilde{k}_{F_0} = F_0 \cdot \frac{K}{f_s}$, on basis of the simple relation that $\frac{K}{2}$ corresponds to $\frac{f_s}{2}$ and every frequency bin index containing a related frequency can be obtained by linear interpolation. Due to the periodicity of the cosine we compute the integer bin index of the first local minimum according to

$$k_{\min} = \left\lceil \frac{\tilde{k}_{F_0}}{2} \right\rceil, \qquad k_{\min} \in \{0, 1, \ldots, K-1\}. \tag{29}$$

The maximum for the high frequencies is found by analyzing whether the highest possible harmonic frequency and thus the corresponding period of the cosine fits into the non-redundant frequency range as limited by $\frac{f_s}{2}$, or not, according to (30) shown at the bottom of the page.

Here, we have to distinguish between two cases: either, the highest depictable harmonic frequency ($F_0 \cdot \lfloor \frac{f_s}{2F_0} \rfloor$) including its falling edge ($+\frac{F_0}{2}$) fits into the non-redundant frequency range ($\leq \frac{f_s}{2}$) or it overlaps ($> \frac{f_s}{2}$). For the former we simply calculate the frequency of the last local minimum at the end of the falling edge of the last harmonic ($F_0 \cdot \lfloor \frac{f_s}{2F_0} \rfloor + \frac{F_0}{2}$) and calculate its corresponding frequency bin index ($\cdot \frac{K}{f_s}$). For the latter, since this frequency would be outside of the non-redundant frequency range, we calculate it for the last but one harmonic ($\lfloor \frac{f_s}{2F_0} \rfloor - 1$), accordingly. For more clarity we refer to the lower panel of Fig. 5, depicting $k_{\min}$, $\tilde{k}_{F_0}$, and also $k_{\max}$.

For all $k < k_{\min}$ and $k > k_{\max}$ the real-valued cosine in the log-spectral domain is discarded and simply to be extended linearly with the slope around $k = k_{\min}$ and $k = k_{\max}$, respectively. The proposed mechanism is *one* possibility to solve the issue quite well already, as a comparison of Fig. 5, upper and lower panel, suggests. Alternatively, different monotonically

$$k_{\max} = \begin{cases} \left\lceil \left(F_0 \cdot \left\lfloor \frac{f_s}{2F_0} \right\rfloor + \frac{F_0}{2} \right) \cdot \frac{K}{f_s} \right\rceil, & \text{for } F_0 \cdot \left\lfloor \frac{f_s}{2F_0} \right\rfloor + \frac{F_0}{2} \leq \frac{f_s}{2} \\ \left\lceil \left(F_0 \cdot \left(\left\lfloor \frac{f_s}{2F_0} \right\rfloor - 1 \right) + \frac{F_0}{2} \right) \cdot \frac{K}{f_s} \right\rceil, & \text{for } F_0 \cdot \left\lfloor \frac{f_s}{2F_0} \right\rfloor + \frac{F_0}{2} > \frac{f_s}{2} \end{cases} \tag{30}$$

https://doi.org/10.24355/dbbs.084-202102121510-0

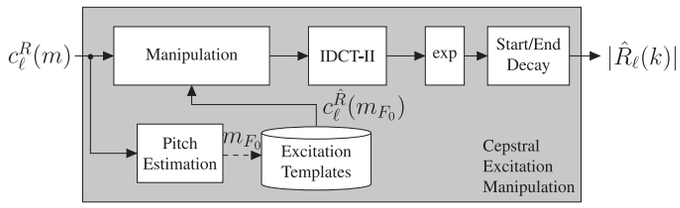


Fig. 6. Block diagram of the **proposed cepstral excitation manipulation** based on **excitation templates**.

increasing and decreasing functions could be applied to model the start and end decay as required.

B. *Cepstral Excitation Templates* ($\mathrm{CEM_{SI}}$, $\mathrm{CEM_{SD}}$)

In the cepstral excitation manipulation as described before, the idealized synthetic excitation carries no specific information except for the location of harmonics and a proper energy coefficient. However, further investigations have shown that a residual signal that is obtained via LPC analysis and subsequently transformed into the cepstral domain holds non-negligible information in the remaining bins. As a consequence, the synthetic excitation lacks attributes of the human vocal chords and lungs, being responsible for the naturalness of the modeled excitation signal. Thus, with the now described alternative approach we aim at modeling these components in more detail by creating excitation templates based on excitation signals originating from LPC analysis. A high-level diagram of the cepstral excitation manipulation is depicted in Fig. 6. We obtain cepstral excitation templates in two different ways, depending on whether they are going to be speaker-independent (SI) or speaker-dependent (SD). In the following we describe a general way which is used for both, SI and SD templates, where for the latter it is just a first stage towards a more speaker-specific modeling. In general, the idea is to have a cepstral excitation template for each detectable pitch bin value $m \in \mathcal{M}_{F_0}$. For this, clean speech training material is analyzed and subsequently the DFT spectrum $S_\ell(k)$ of each frame ℓ is separated into spectral envelope and excitation signal (see Fig. 1 and assume $\bar{Y}_\ell(k) = S_\ell(k)$). The DCT-II is employed to transform the excitation signal into the cepstral domain and the pitch bin index m_{F_0} is estimated as explained in Section III-C. Accordingly, we collect the set

$$\mathcal{C}_{m_{F_0}} = \left\{ \mathbf{c}_\ell^R | c_\ell^R(m_{F_0}) \geq c_\ell^R(\mu)\ \forall \mu \in \mathcal{M}_{F_0} \right\} \tag{31}$$

of all cepstral vectors in the training material belonging to each particular pitch bin index $m_{F_0} \in \mathcal{M}_{F_0}$, with $\mathbf{c}_\ell^R = \left(c_\ell^R(0), \ldots, c_\ell^R(m), \ldots, c_\ell^R(K-1)\right)$. Now, the cepstral representation allows us to average per bin over all cepstral excitations in a given set and to obtain a representative excitation template for each pitch bin index m_{F_0} as

$$\bar{\mathbf{c}}^R(m_{F_0}) = \frac{1}{|\mathcal{C}_{m_{F_0}}|} \sum_{\mathbf{c}^R \in \mathcal{C}_{m_{F_0}}} \mathbf{c}^R, \quad \forall m_{F_0} : |\mathcal{C}_{m_{F_0}}| > 0, \tag{32}$$

where $|\cdot|$ is the cardinality of a certain set. If a set is empty, the codebook entry is assigned an all-zero vector. Furthermore, we drop the frame index ℓ as it is only required during the collection of the training material in (31).

The templates can be obtained either in an SI or an SD fashion, only depending on the training data. We propose to use the following adaptation scheme to create SD templates on basis of SI templates. At first, we separately obtain the SI templates $\bar{\mathbf{c}}^R(m_{F_0})$ and preliminary SD templates $\check{\mathbf{c}}^R(m_{F_0})$ stemming from much less data of the target speaker, both according to (31) and (32), differing only in the training material. The actual adaptation is a weighted mixture of both, SI and SD templates for each given pitch bin index m_{F_0} as

$$\tilde{\mathbf{c}}^R(m_{F_0}) = \beta(m_{F_0}) \cdot \bar{\mathbf{c}}^R(m_{F_0}) + (1 - \beta(m_{F_0})) \cdot \check{\mathbf{c}}^R(m_{F_0}) \tag{33}$$

with

$$\beta(m_{F_0}) = \frac{|\bar{\mathcal{C}}_{m_{F_0}}|}{|\bar{\mathcal{C}}_{m_{F_0}}| + \delta \cdot |\check{\mathcal{C}}_{m_{F_0}}|} \tag{34}$$

where $\delta \geq 1$ allows to compensate for the typical lack of SD training material and to artificially emphasize the SD material.

Having obtained and stored the cepstral excitation templates, their application is very similar to the scheme in Section IV-A. After having detected the pitch bin index m_{F_0} according to (25), the SD (33) or SI (32) cepstral excitation template addressed by m_{F_0} is taken. Here, e.g., for SI templates:

$$\hat{c_\ell^R}(m) = \bar{c}^R(m_{F_0}, m) \qquad \forall m \notin \{0, m_{F_0}\}. \tag{35}$$

The subsequent manipulations from (26) to (28) are applied as before in order to obtain a level consistent with the preliminary denoised signal where (27) is replaced by (35). The proposed start and end decay from (29) and (30) can optionally be applied to compensate for aberrations due to noise in the training material.

If an empty template (originating from $|\mathcal{C}_{m_{F_0}}| = 0$ during training) has been selected by the detected pitch bin index m_{F_0}, we do *not* apply the manipulations from (26) to (28). Instead, we continue with the all-zero cepstral template $\bar{\mathbf{c}}^R(m_{F_0})$ which results in a flat spectrum with unity amplitude. Thus, in such a situation of uncertainty, we do not harm the signal nor do we necessarily enhance it since the clean speech estimate $|\hat{S}_\ell(k)|$ then reduces to solely the envelope $|H_\ell(k)|$. Alternatively, one could also employ the idealized approach in such cases or learn missing templates in an adaptive manner. We comment on the amount of empty templates and their selection frequency during test at the end of Section V-D.

V. Experimental Evaluation

We embed the proposed and the baseline *a priori* SNR estimators in a common noise reduction algorithm to evaluate their performance and analyze their behavior in four different noise types, six different SNR conditions and with two commonly employed spectral weighting rules. Four different quality measures are utilized to compare the different approaches.

A. *Experimental Setup*

Throughout the whole experimental section of this contribution we employ a sample rate $f_s = 8$ kHz with a frame size of 32 ms, corresponding to $K = 256$ samples, and a 50% frame

https://doi.org/10.24355/dbbs.084-202102121510-0

shift by 128 samples. As analysis and overlap-add synthesis window we utilize a periodic square root Hann window and for the source-filter decomposition we compute $N = 10$ LPC coefficients.

The NTT super wideband database [32] is used as a basis and thus downsampled to 8 kHz. We only use the American and British English sets which consist of eight and six speakers, respectively, where each set offers an equal number of speakers per gender. The database comes with 120 utterances for each American English and 100 for each British English speaker. Thus, we decided to artificially decrease the amount to 100 files for American English speakers by random picking, amounting to a total of 14 speakers and 1400 utterances. Next, we use 80% of each speaker's material for training and the remaining 20% for SI and SD testing. For our SI experiments, we decided to use a leave-one-out method to increase the amount of training material. For this, we generate a training set for each speaker separately containing the training material of the 13 other speakers consisting of $13 \times 80 = 1040$ utterances. The training material for the SD adaptation is represented by the 80 utterances of each speaker which have been left out during the SI training. The training itself of the SI and SD templates is conducted according to Section IV-B with applied start and end decay.

The four different noise types are taken from the ETSI [33] database and represent road, car, office, and pub noise. Each segment used to generate the microphone signal is randomly extracted matching the length of the clean speech file. We process the files at six different SNR conditions ranging from -5 dB up to 20 dB in steps of 5 dB. The level of the clean speech and the noise is measured by the active speech level and the root-mean-square level, respectively, according to ITU-T P.56 [34] and both adjusted also according to P.56 prior to superposition. In total we process $14 \times 20 \times 4 \times 6 = 6720$ files for each *a priori* SNR estimator under test. Please note that for the SD experiments, we switch the SD templates corresponding to each speaker being processed, accordingly.

The evaluation of the different *a priori* SNR estimators is placed in a common noise reduction system with MS noise power estimation, one of the *a priori* SNR estimators under test (DD, HRNR, CTS, CEM_{ID}, CEM_{SI}, CEM_{SD}) and the two spectral weighting rules (MMSE-LSA and SG-jMAP) used to calculate the final gains $G_\ell(k)$ which are limited to $G_{\min} = -15$ dB.

The DD approach is tuned with optimal parameters[3] adopted from [35] for each of the weighting rules.

For the HRNR approach a preliminary noise reduction is required for which we also use the MS noise power estimation, DD *a priori* SNR estimation with $\beta_{\text{DD}} = 0.985$ and $\xi_{\min} = -15$ dB as it is just an intermediate step. Furthermore, $G_\ell^{\text{DD}}(k)$ and $G_\ell^{\text{TSNR}}(k)$ are calculated using the WF as proposed and we follow the author's suggestion and utilize $\alpha_\ell(k) = G_\ell^{\text{TSNR}}(k)$ as weights for the mixing in (16) .

The CTS implementation was kindly provided by the authors and thus the parameters left as originally initialized.

Our three proposed estimators CEM_{ID}, CEM_{SI}, and CEM_{SD}, share the same preliminary noise reduction with the HRNR approach except for the weighting rule being MMSE-LSA (instead of the WF) as mentioned in Section III-A. The overestimation factor for (26) is empirically determined and set to $\alpha_\ell(m_{F_0}) = 2$. The required parameter to compensate the lack of speaker-dependent data is found for this particular training set with $\delta = 30$.

B. Quality Measures

To measure the quality of our proposed *a priori* SNR estimator in two example noise reduction contexts, we employ the so-called white-box approach [36], i.e., we calculate the gains $G_\ell(k)$ and subsequently apply it not only to the microphone signal $Y_\ell(k)$ in order to obtain the enhanced signal, but also to the clean speech component $S_\ell(k)$ and the noise component $D_\ell(k)$, separately. The obtained components after IDFT and overlap-add are called the *filtered* clean speech component $\tilde{s}(n)$ and the *filtered* noise component $\tilde{d}(n)$, respectively, which is applicable by assuming (2). The measures are operating *not* on the enhanced signal $\hat{s}(n)$, but only on the filtered and unfiltered *components* with the latter as a reference.

The segmental noise attenuation (NA) [37] is calculated as

$$\text{NA}_{\text{seg}} = 10 \log_{10} \left[\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{NA}(\ell) \right], \tag{36}$$

with

$$\text{NA}(\ell) = \frac{\sum_{\nu=0}^{N-1} d(\nu + \ell N)^2}{\sum_{\nu=0}^{N-1} \tilde{d}(\nu + \ell N + \Delta)^2},$$

where ℓ defines a segment of length $N = 256$ samples, Δ is compensating the sample delay of the filtered signal, and $\frac{1}{|\mathcal{L}|}$ is a normalization factor since $|\mathcal{L}|$ is the cardinality of the set $\mathcal{L}$, containing all frames. The segmental NA depicts the average of a local frame-wise ratio of the noise component and the corresponding filtered noise component and is sought to be high.

Different from that we define a global measure

$$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}, \tag{37}$$

where SNR_{in} is the SNR of the clean speech and noise component measured according to ITU P.56 [34], and SNR_{out} correspondingly for the *filtered* components. This measure gives a more general information of the achieved noise suppression over the whole file compared to the segmental NA. A positive ΔSNR indicates an improved SNR after processing.

Please note that the segmental NA and the ΔSNR are not directly related due to their different scopes (local and global) and have to be interpreted separately, as a high segmental NA is not necessarily indicating great SNR improvement and vice versa.

[3]Optimal parameters for DD *a priori* SNR estimation for the two weighting rules:

$$\text{MMSE-LSA: } \beta_{\text{DD}} = 0.975, \xi_{\text{min}} = -15 \text{ dB}$$

$$\text{SG-jMAP: } \beta_{\text{DD}} = 0.993, \xi_{\text{min}} = -14 \text{ dB}.$$

https://doi.org/10.24355/dbbs.084-202102121510-0

TABLE I
DETAILED EVALUATION OF SEGMENTAL NA, ΔSNR, PESQ MOS-LQO, AND SEGMENTAL SSDR, FOR THE FOUR DIFFERENT NOISE TYPES, FIVE SNR CONDITIONS, THE BASELINES VS. THE PROPOSED *A Priori* SNR ESTIMATORS, AND THE **MMSE-LSA** SPECTRAL WEIGHTING RULE

		NA_{seg} [dB]						ΔSNR [dB]						PESQ MOS-LQO						$SSDR_{seg}$ [dB]					
	SNR [dB]	-5	0	5	10	15	mean	-5	0	5	10	15	mean	-5	0	5	10	15	mean	-5	0	5	10	15	mean
ROAD	DD	12.43	12.24	12.02	11.80	11.60	11.91	9.96	10.21	10.12	9.81	9.42	9.74	3.63	3.83	3.98	4.10	4.21	4.01	10.69	14.21	18.19	22.09	25.24	19.63
	HRNR	**14.00**	**13.66**	13.23	12.83	12.46	13.05	11.34	11.57	10.94	9.86	8.51	9.85	3.15	3.39	3.57	3.75	3.92	3.64	6.46	9.19	12.55	16.22	20.02	14.65
	CTS	12.15	11.98	11.86	11.77	11.67	11.83	8.25	8.73	8.69	8.40	7.94	8.23	**3.68**	**3.89**	**4.02**	**4.13**	4.21	**4.04**	9.54	13.47	17.69	21.67	24.87	19.06
	CEM_{ID}	13.81	13.53	13.24	12.94	12.65	13.08	11.83	11.92	11.52	10.75	9.66	10.66	3.60	3.80	3.97	4.11	**4.22**	4.00	9.56	13.20	17.38	21.55	25.00	19.00
	CEM_{SI}	13.80	13.55	**13.29**	**13.03**	12.78	**13.16**	12.05	12.08	11.73	11.14	10.39	11.14	3.60	3.79	3.95	4.09	4.21	3.99	10.51	14.32	18.46	22.24	25.22	19.66
	CEM_{SD}	13.79	13.54	13.28	**13.03**	**12.79**	**13.16**	**12.06**	**12.09**	**11.74**	**11.17**	**10.45**	**11.18**	3.60	3.79	3.95	4.09	4.21	3.99	**10.71**	**14.51**	**18.60**	**22.34**	**25.28**	**19.78**
CAR	DD	11.82	11.80	11.74	11.66	11.54	11.65	7.04	7.28	7.35	7.38	7.36	7.29	**4.19**	4.28	4.34	4.39	4.43	**4.35**	20.54	23.90	26.53	28.17	**28.99**	26.25
	HRNR	12.86	12.39	11.96	11.61	11.29	11.85	6.99	6.68	5.72	4.65	3.78	5.15	3.78	3.96	4.13	4.25	4.32	4.13	13.41	18.03	22.22	25.09	27.08	22.36
	CTS	12.23	12.27	12.25	12.20	12.12	12.18	6.99	6.91	6.72	6.49	6.25	6.56	**4.19**	4.28	4.34	4.39	4.42	4.34	20.09	23.47	26.14	27.94	28.93	25.99
	CEM_{ID}	14.54	14.41	14.26	14.05	13.78	14.08	14.22	14.22	14.07	13.73	13.18	13.62	4.17	**4.29**	**4.35**	**4.40**	**4.44**	**4.35**	20.66	24.27	**26.74**	**28.19**	28.97	**26.37**
	CEM_{SI}	**14.60**	**14.51**	**14.41**	14.27	14.09	14.29	**14.40**	14.41	**14.34**	14.16	13.85	14.08	4.15	4.27	4.34	**4.40**	4.43	4.34	21.53	24.32	26.40	27.78	28.61	26.29
	CEM_{SD}	**14.60**	**14.51**	**14.41**	**14.28**	**14.11**	**14.30**	**14.40**	**14.42**	**14.34**	**14.17**	**13.88**	**14.10**	4.16	4.27	4.34	**4.40**	4.43	4.34	**21.64**	**24.40**	26.47	27.84	28.67	26.36
OFFICE	DD	8.11	7.99	7.87	7.76	7.66	7.83	1.47	2.03	2.32	2.44	2.47	2.20	3.56	3.78	3.94	4.07	**4.18**	3.97	**11.36**	**14.85**	**18.60**	**22.21**	25.17	**19.92**
	HRNR	9.71	9.43	9.20	9.03	8.90	9.18	0.86	1.91	2.48	2.68	2.64	2.18	3.01	3.28	3.50	3.68	3.85	3.55	6.21	8.56	11.78	15.65	19.63	14.17
	CTS	9.42	9.28	9.20	9.14	9.11	9.21	1.44	2.27	2.67	2.85	2.88	2.49	**3.61**	**3.83**	**3.98**	**4.09**	4.17	**3.99**	10.27	13.85	17.81	21.57	24.67	19.18
	CEM_{ID}	**10.21**	**9.98**	9.79	9.64	9.52	9.76	3.02	3.74	4.06	4.11	4.01	3.80	3.48	3.74	3.92	4.07	**4.18**	3.94	9.52	13.20	17.57	21.89	**25.34**	19.18
	CEM_{SI}	10.16	9.97	9.81	9.67	9.55	**9.77**	3.31	4.02	4.32	4.38	4.29	4.08	3.47	3.73	3.91	4.05	**4.18**	3.94	9.91	13.81	18.02	21.93	25.00	19.30
	CEM_{SD}	10.16	9.97	**9.82**	**9.68**	**9.56**	**9.77**	**3.37**	**4.05**	**4.35**	**4.40**	**4.32**	**4.11**	3.48	3.74	3.91	4.06	**4.18**	3.94	10.11	14.01	18.19	22.04	25.06	19.43
PUB	DD	7.72	7.59	7.46	7.37	7.31	7.46	1.49	2.42	2.90	3.06	3.04	2.64	3.16	3.51	3.76	3.94	4.07	3.77	**8.19**	**11.32**	**14.83**	**18.62**	22.33	**16.79**
	HRNR	**10.29**	**10.06**	**9.83**	**9.66**	**9.55**	**9.81**	1.22	3.14	**4.31**	**4.83**	**4.86**	**3.84**	2.65	2.85	3.15	3.44	3.67	3.27	4.29	6.18	9.01	12.80	16.71	11.58
	CTS	9.44	9.25	9.11	9.04	9.02	9.15	1.55	2.90	3.60	3.86	3.84	3.23	**3.26**	**3.58**	**3.80**	**3.96**	**4.08**	**3.81**	6.80	9.86	13.74	17.98	22.05	15.95
	CEM_{ID}	9.39	9.12	8.89	8.74	8.62	8.88	1.37	3.04	3.75	3.88	3.72	3.20	2.90	3.33	3.70	3.92	4.07	3.68	5.62	9.05	13.55	18.36	**22.81**	15.92
	CEM_{SI}	9.22	9.00	8.80	8.66	8.54	8.78	1.81	3.35	3.99	4.09	3.92	3.47	2.91	3.35	3.69	3.90	4.06	3.68	6.14	9.70	14.06	18.50	22.57	16.12
	CEM_{SD}	9.21	8.99	8.80	8.66	8.54	8.77	**1.91**	**3.41**	4.03	4.12	3.95	3.52	2.92	3.36	3.69	3.90	4.06	3.69	6.29	9.89	14.24	18.61	22.62	16.23
Means	DD	10.02	9.90	9.77	9.65	9.53	9.71	4.99	5.49	5.67	5.67	5.57	5.47	3.63	3.85	4.01	4.13	4.22	4.02	**12.69**	**16.07**	**19.53**	**22.77**	25.43	**20.65**
	HRNR	11.72	11.38	11.06	10.78	10.55	10.97	5.10	5.83	5.86	5.51	4.95	5.25	3.15	3.37	3.59	3.78	3.94	3.65	7.59	10.49	13.89	17.44	20.86	15.69
	CTS	10.81	10.70	10.60	10.54	10.48	10.59	4.56	5.20	5.42	5.40	5.23	5.13	**3.68**	**3.89**	**4.04**	**4.14**	4.22	**4.04**	11.68	15.16	18.85	22.29	25.13	20.05
	CEM_{ID}	**11.99**	**11.76**	11.55	11.34	11.14	11.45	7.61	8.23	8.35	8.12	7.64	7.82	3.54	3.79	3.99	4.12	**4.23**	4.00	11.34	14.93	18.81	22.50	**25.53**	20.12
	CEM_{SI}	11.95	**11.76**	**11.58**	**11.41**	11.24	**11.50**	7.89	8.47	8.59	8.44	8.11	8.19	3.53	3.78	3.97	4.11	4.22	3.99	12.02	15.54	19.24	22.61	25.35	20.34
	CEM_{SD}	11.94	11.75	**11.58**	**11.41**	**11.25**	**11.50**	**7.94**	**8.49**	**8.61**	**8.47**	**8.15**	**8.23**	3.54	3.79	3.98	4.11	4.22	3.99	12.19	15.70	19.38	22.71	25.41	20.45

The first measure to assess the quality of the *filtered* clean speech component is the segmental speech-to-speech-distortion ratio (SSDR) [37] calculated as

$$\mathrm{SSDR}_{\mathrm{seg}} = \frac{1}{|\mathcal{L}_1|} \sum_{\ell \in \mathcal{L}_1} \mathrm{SSDR}(\ell) \tag{38}$$

where $\mathcal{L}_1$ depicts the set of speech active frames obtained by a simple energy threshold-based voice activity detection operating on the clean speech signal $s(n)$. Additionally, $\mathrm{SSDR}(\ell)$ is limited to values between -10 dB and 30 dB by

$$\mathrm{SSDR}(\ell) = \max\left\{\min\left\{\mathrm{SSDR}'(\ell), R_{\max}\right\}, R_{\min}\right\}.$$

The actual ratio necessary for computation is obtained by

$$\mathrm{SSDR}'(\ell) = 10\log_{10}\left[\frac{\sum_{\nu=0}^{N-1} s(\nu+\ell N)^2}{\sum_{\nu=0}^{N-1} e(\nu+\ell N)^2}\right]$$

where the speech distortion is

$$e(\nu+\ell N) = \tilde{s}(\nu+\ell N+\Delta) - s(\nu+\ell N).$$

A high segmental SSDR indicates low speech distortion and thus good preservation of the speech component.

Furthermore, we employ the PESQ mean opinion score (MOS-LQO) [38] to obtain another measure for the quality of the *filtered* clean speech component. Please note that in line with P.1100 [39, Sec. 8] we do not utilize the enhanced speech $\hat{s}(n)$ in the PESQ measure but the separately processed speech component $\tilde{s}(n)$ as PESQ has not been validated for potential artifacts caused by noise reduction algorithms. We aim at being more compliant to P.862 [38] by doing so.

C. Experimental Results: Details

We provide a detailed evaluation for both the MMSE-LSA and the SG-jMAP spectral weighting rule in the following Tables I and II. Each table depicts the four quality measures for all of the noise types in the SNR conditions from -5 dB to 15 dB averaged over the whole test set, where the best scores are highlighted in boldface. For the computation of the mean over SNRs also the 20 dB SNR condition has been included, which is, however, left out as separate column simply due to space restrictions. This allows for a very extensive analysis of the tested *a priori* SNR estimators for each condition separately.

In Table I the performance results for the MMSE-LSA spectral weighting rule are shown. In terms of noise suppression (measures $\mathrm{NA}_{\mathrm{seg}}$ and ΔSNR) the CEM approaches clearly show the strongest performance for each SNR condition, averaged over all four noise types. Both the DD and the CTS baselines show poor performance, and have only few SNR/noise type conditions with convincing performance. The HRNR approach is on average in many cases the best of the baseline approaches w.r.t. noise suppression, showing particularly good performance in pub noise (in $\mathrm{NA}_{\mathrm{seg}}$, and for medium to high SNRs also best

https://doi.org/10.24355/dbbs.084-202102121510-0

TABLE II
DETAILED EVALUATION OF SEGMENTAL NA, ΔSNR, PESQ MOS-LQO, AND SEGMENTAL SSDR, FOR THE FOUR DIFFERENT NOISE TYPES, FIVE SNR CONDITIONS, THE BASELINES VS. THE PROPOSED *A Priori* SNR ESTIMATORS, AND THE **SG-JMAP** SPECTRAL WEIGHTING RULE

		NA_{seg} [dB]						ΔSNR [dB]						PESQ MOS-LQO						$SSDR_{seg}$ [dB]					
	SNR [dB]	-5	0	5	10	15	mean	-5	0	5	10	15	mean	-5	0	5	10	15	mean	-5	0	5	10	15	mean
ROAD	DD	13.25	13.06	12.84	12.64	12.45	12.75	10.61	10.80	10.66	10.39	9.91	10.27	3.63	3.75	3.90	4.05	4.18	3.97	11.22	15.08	19.31	23.27	26.18	20.50
	HRNR	13.36	13.08	12.73	12.40	12.12	12.59	10.44	10.23	9.36	8.86	7.59	8.72	3.41	3.58	3.80	4.01	4.17	3.87	11.36	15.65	20.22	24.23	26.91	21.14
	CTS	12.57	12.40	12.25	12.16	12.07	12.24	8.98	9.19	9.05	8.82	8.34	8.69	**3.76**	**3.86**	**4.00**	**4.14**	**4.24**	**4.05**	**11.85**	**16.22**	**20.66**	**24.44**	**27.03**	**21.45**
	CEM_{ID}	**13.78**	13.53	13.26	13.00	12.76	13.14	11.98	12.03	11.71	11.04	10.18	11.01	3.69	3.82	3.97	4.11	4.23	4.02	11.06	15.21	19.73	23.77	26.67	20.80
	CEM_{SI}	**13.78**	**13.55**	**13.31**	13.07	12.86	13.20	**12.11**	**12.14**	**11.87**	11.34	10.70	11.35	3.68	3.80	3.95	4.10	4.23	4.01	11.63	15.83	20.19	23.93	26.61	21.06
	CEM_{SD}	13.77	**13.55**	**13.31**	**13.08**	**12.87**	**13.21**	**12.11**	**12.14**	**11.87**	**11.35**	**10.74**	**11.37**	3.68	3.80	3.95	4.10	4.23	4.02	11.76	15.94	20.27	23.99	26.64	21.13
CAR	DD	13.00	12.97	12.89	12.82	12.73	12.84	6.66	6.90	6.94	7.01	6.95	6.89	4.15	4.25	4.33	4.39	4.43	4.33	21.66	24.94	27.34	28.60	29.19	26.87
	HRNR	12.94	12.55	12.09	11.72	11.40	11.98	5.91	5.57	5.05	3.93	3.18	4.35	3.99	4.15	4.27	4.34	4.39	4.26	22.26	25.53	27.76	28.90	**29.46**	27.26
	CTS	12.56	12.59	12.56	12.51	12.45	12.51	6.21	6.29	6.19	6.12	5.97	6.10	**4.21**	**4.30**	**4.37**	**4.41**	**4.44**	**4.36**	22.73	25.88	**27.94**	**29.00**	**29.46**	**27.44**
	CEM_{ID}	14.65	14.54	14.42	14.26	14.06	14.29	14.12	14.12	14.01	13.78	13.39	13.70	**4.21**	**4.30**	4.36	**4.41**	**4.44**	**4.36**	22.83	**25.96**	27.87	28.87	29.38	27.42
	CEM_{SI}	**14.69**	**14.61**	**14.52**	**14.41**	14.28	**14.44**	**14.36**	**14.39**	**14.32**	14.18	13.96	**14.14**	4.19	4.29	4.36	4.40	4.43	4.35	22.96	25.74	27.59	28.64	29.20	27.27
	CEM_{SD}	**14.69**	**14.61**	**14.52**	**14.41**	**14.29**	**14.44**	**14.36**	14.38	**14.32**	**14.19**	**13.97**	**14.14**	4.19	4.29	4.36	4.40	4.43	4.35	**23.01**	25.79	27.63	28.66	29.22	27.30
OFFICE	DD	8.45	8.35	8.26	8.19	8.16	8.26	1.49	1.96	2.21	2.31	2.33	2.10	3.57	3.75	3.89	4.02	4.15	3.94	12.02	15.97	19.99	23.60	26.29	20.99
	HRNR	8.16	7.98	7.82	7.72	7.68	7.85	1.25	1.60	1.51	1.56	1.17	1.32	3.38	3.63	3.81	3.98	4.13	3.86	11.95	16.53	21.05	**24.76**	**27.23**	21.70
	CTS	8.82	8.71	8.65	8.62	8.65	8.70	1.44	1.98	2.24	2.32	2.36	2.09	**3.73**	**3.87**	**3.99**	**4.10**	**4.21**	**4.03**	**12.57**	**16.90**	**21.18**	24.70	27.14	**21.85**
	CEM_{ID}	9.72	9.55	9.41	9.32	9.27	9.42	2.87	3.44	3.68	3.72	3.68	3.50	3.65	3.83	3.97	**4.10**	**4.21**	4.01	11.47	15.92	20.63	24.58	27.20	21.40
	CEM_{SI}	**9.75**	9.60	**9.49**	**9.40**	9.35	**9.49**	3.05	3.62	3.86	3.92	3.89	3.69	3.63	3.81	3.96	4.09	4.20	4.00	11.61	16.06	20.54	24.27	26.86	21.29
	CEM_{SD}	**9.75**	**9.61**	**9.49**	**9.40**	**9.36**	**9.49**	**3.08**	**3.64**	**3.87**	**3.93**	**3.90**	**3.71**	3.64	3.81	3.96	4.09	4.20	4.00	11.76	16.19	20.63	24.32	26.87	21.36
PUB	DD	7.85	7.74	7.65	7.61	7.60	7.68	1.35	2.31	2.75	2.91	2.93	2.52	3.26	3.54	3.72	3.87	4.02	3.76	**8.47**	12.10	16.03	20.07	23.83	17.86
	HRNR	7.67	7.53	7.44	7.40	7.39	7.48	1.20	2.09	2.35	2.46	2.48	2.15	3.01	3.36	3.64	3.83	4.00	3.66	7.97	12.27	16.94	21.49	**25.35**	18.64
	CTS	8.54	8.40	**8.32**	**8.30**	**8.34**	**8.39**	1.25	2.39	2.90	3.06	3.08	2.62	**3.51**	**3.71**	**3.85**	**3.98**	**4.10**	**3.89**	8.38	**12.58**	**17.21**	**21.63**	25.30	**18.80**
	CEM_{ID}	**8.66**	**8.46**	8.31	8.23	8.19	8.34	1.30	2.62	3.12	3.22	3.14	2.73	3.24	3.59	3.80	3.95	**4.10**	3.82	7.10	11.52	16.57	21.40	25.24	18.25
	CEM_{SI}	8.59	8.42	8.29	8.22	8.18	8.31	1.51	2.77	3.26	3.36	3.28	2.89	3.24	3.58	3.78	3.94	4.09	3.80	7.33	11.72	16.56	21.13	24.90	18.17
	CEM_{SD}	8.59	8.42	8.29	8.22	8.18	8.31	**1.56**	**2.80**	**3.28**	**3.38**	**3.30**	**2.91**	3.24	3.58	3.78	3.94	4.09	3.81	7.45	11.84	16.63	21.15	24.90	18.23
Means	DD	10.64	10.53	10.41	10.31	10.24	10.38	5.02	5.49	5.64	5.66	5.53	5.44	3.65	3.82	3.96	4.08	4.19	4.00	13.34	17.02	20.67	23.89	26.37	21.55
	HRNR	10.53	10.28	10.02	9.81	9.65	9.97	4.70	4.87	4.57	4.20	3.61	4.13	3.45	3.68	3.88	4.04	4.17	3.91	13.38	17.49	21.49	24.85	**27.24**	22.19
	CTS	10.62	10.52	10.44	10.40	10.38	10.46	4.47	4.96	5.09	5.08	4.94	4.87	**3.80**	**3.93**	**4.05**	**4.16**	**4.25**	**4.09**	**13.88**	**17.90**	**21.75**	**24.94**	27.23	**22.39**
	CEM_{ID}	**11.70**	11.52	11.35	11.20	11.07	11.30	7.57	8.05	8.13	7.94	7.60	7.73	3.70	3.89	4.03	4.14	4.24	4.05	13.12	17.15	21.20	24.66	27.12	21.97
	CEM_{SI}	**11.70**	**11.55**	**11.40**	11.27	11.16	**11.36**	7.76	8.23	8.33	8.20	7.96	8.02	3.68	3.87	4.01	4.13	4.24	4.04	13.38	17.34	21.22	24.49	26.89	21.95
	CEM_{SD}	**11.70**	**11.55**	**11.40**	**11.28**	**11.17**	**11.36**	**7.78**	**8.24**	**8.34**	**8.21**	**7.98**	**8.03**	3.69	3.87	4.01	4.13	4.24	4.04	13.49	17.44	21.29	24.53	26.91	22.01

in ΔSNR). On the contrary, in the SNR $= -5$ dB condition, HRNR's ΔSNR in pub noise is worst among all schemes, while it yields the best NA_{seg} in road noise at that SNR. The proposed CEM schemes are much more consistent in terms of NA_{seg} and ΔSNR over SNR conditions and noises: The speaker-dependent CEM_{SD} is best in all cases, except only for NA_{seg} in very low SNR, where CEM_{ID} is slightly ahead.

In terms of the speech component quality (PESQ, $SSDR_{seg}$) the picture is partly different: On average over the noise types CTS performs best in PESQ in most SNRs, being ahead up to 0.15 MOS points vs. the worst CEM approach. Interestingly, however, in car noise, CEM_{ID} is slightly better than CTS in most SNR conditions. The classical DD approach performs on a par with other approaches with regard to the PESQ metric, while HRNR consistently fails to provide an acceptable speech component quality, both in PESQ and $SSDR_{seg}$. Surprisingly, DD delivers very good $SSDR_{seg}$ performance in office and pub noise, while the CEM approaches perform best in car and road noise.

In Table II the performance results for the SG-jMAP spectral weighting rule are shown. In terms of noise suppression (measures NA_{seg} and ΔSNR) the CEM approaches, especially the speaker-dependent variant CEM_{SD}, clearly perform best in each SNR condition, averaged over all four noise types. Both the DD and here the HRNR baselines show poor performance, and have no single SNR/noise type condition with convincing NA_{seg} performance. Considering ΔSNR, *none* of the three baselines has a single SNR/noise type condition with superior performance. The CTS approach is on average in most cases the best of the baseline approaches w.r.t. NA_{seg}, showing particularly good performance in pub noise for medium to high SNRs. Interestingly, for the SG-jMAP, the DD approach is on average the best baseline w.r.t. ΔSNR, showing that sophisticated spectral weighting rules are able to heal some shortcomings of earlier processing stages such as the SNR estimation. The proposed CEM schemes are much more consistent in terms of NA_{seg} and ΔSNR over SNR conditions and noises: The speaker-dependent CEM_{SD} is best in all cases.

In terms of the speech component quality (PESQ, $SSDR_{seg}$) the picture again is partly different: On average over the noise types, CTS performs best in PESQ for all SNRs, being ahead up to 0.12 MOS points vs. the worst CEM approach. In car noise, however, CEM_{ID} is on a par with CTS in most SNR conditions. On average, the DD approach performs quite well in PESQ, while HRNR consistently settles for the worst score. The $SSDR_{seg}$ is also mostly in favor of the CTS approach, and opposite to PESQ, the HRNR approach is found to be slightly ahead of the DD estimator on average.

Please note that the advantage of a speaker-dependent approach vs. all other speaker-independent approaches is of course somehow expected, yet CEM_{SD} is only slightly ahead of our speaker-independent method CEM_{SI}.

We can summarize for both weighting rules, that on average over the noise types the CEM approaches perform best in

https://doi.org/10.24355/dbbs.084-202102121510-0

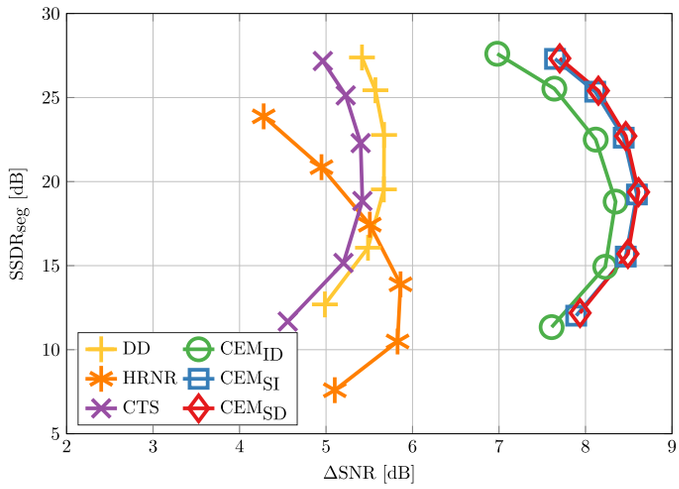


Fig. 7. Segmental SSDR and **ΔSNR** averaged over the four different noise types for the different *a priori* SNR estimators under test with the **MMSE-LSA** spectral weighting rule.

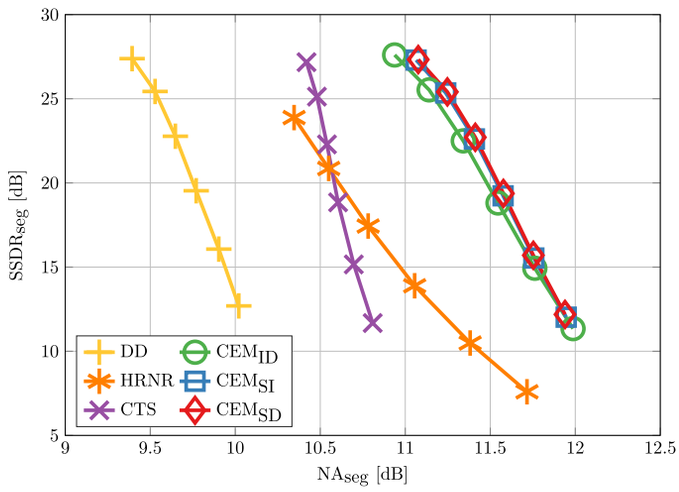


Fig. 9. Segmental SSDR and **segmental NA** averaged over the four different noise types for the different *a priori* SNR estimators under test with the **MMSE-LSA** spectral weighting rule.

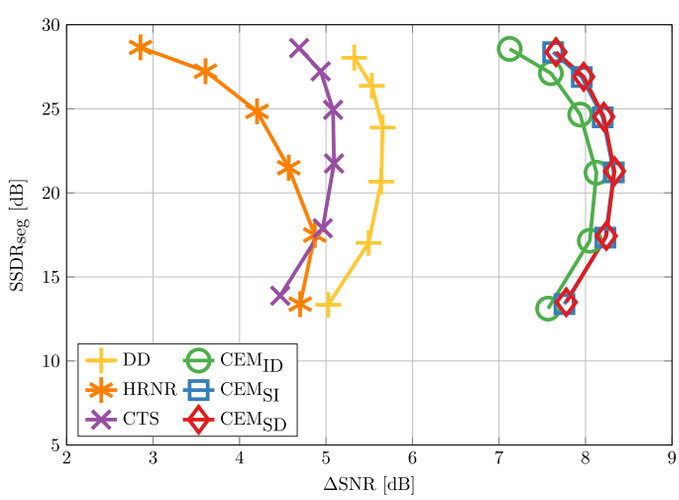


Fig. 8. Segmental SSDR and **ΔSNR** averaged over the four different noise types for the different *a priori* SNR estimators under test with the **SG-jMAP** spectral weighting rule.

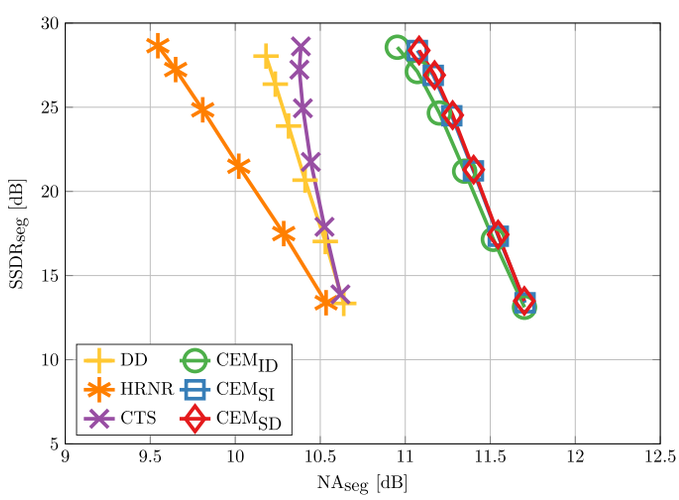


Fig. 10. Segmental SSDR and **segmental NA** averaged over the four different noise types for the different *a priori* SNR estimators under test with the **SG-jMAP** spectral weighting rule.

terms of NA$_{\text{seg}}$ and ΔSNR, and almost on a par with the best performing method w.r.t. PESQ and SSDR$_{\text{seg}}$. The baselines all show an imbalanced performance being inferior in one of the two main categories: For both weighting rules, they are inferior w.r.t. NA$_{\text{seg}}$ and ΔSNR (DD, HRNR and CTS). For the MMSE-LSA weighting rule, HRNR shows poor performance w.r.t. PESQ and SSDR$_{\text{seg}}$, where for the SG-jMAP weighting rule HRNR performs poorly w.r.t. PESQ, while DD is only slightly inferior in SSDR$_{\text{seg}}$.

D. Discussion

To enable further analysis of the results, we plot the SSDR$_{\text{seg}}$ for the two spectral weighting rules over the ΔSNR (Figs. 7 and 8) and the NA$_{\text{seg}}$ (Figs. 9 and 10), respectively. The plots simplify the interpretation on a more global level compared to the tables as only two dimensions are considered at a time. Each plot is a visualization of the mean section from the corresponding table. In each figure the different estimators are specifiable by their respective marker, being $+$ for DD, $*$ for HRNR, and $\times$ for CTS, the three illustrating the baseline algorithms. The proposed techniques are distinguishable by $\circ$ for the idealized synthetic approach CEM$_{\text{ID}}$, $\square$ for the template-based variant CEM$_{\text{SI}}$, and finally $\Diamond$ for the speaker-dependent template-based implementation CEM$_{\text{SD}}$. Each marker depicts one of the SNR conditions from -5 dB to 20 dB in steps of 5 dB, where the lowest corresponds to the worst and the highest to the best condition, respectively. The further a marker is located to the top right hand corner of each plot, the better is the performance. The range and scaling of each axis showing its respective measure is the same to ensure comparability across the two spectral weighting rules.

The MMSE-LSA estimator is depicted in Fig. 7 showing how close DD and CTS are. The HRNR approach exhibits quite an unbalanced behavior as ΔSNR performance is similar to the

https://doi.org/10.24355/dbbs.084-202102121510-0

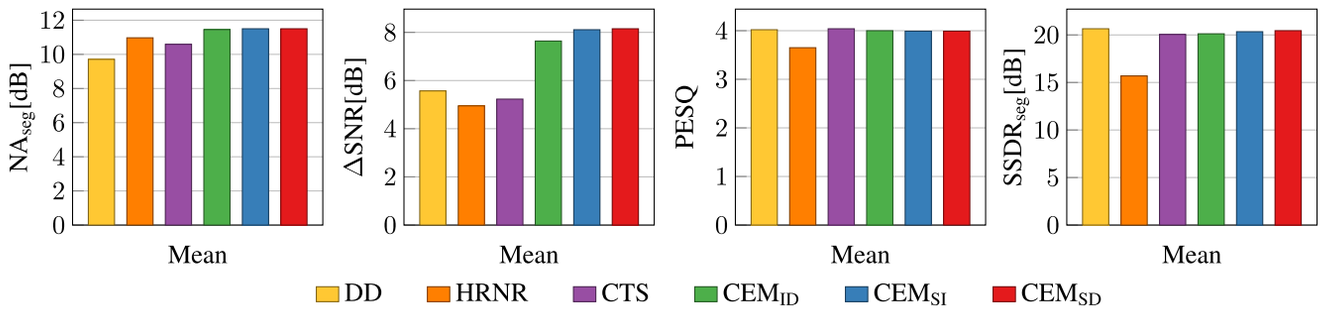


Fig. 11. Evaluation of segmental NA, ΔSNR, PESQ MOS-LQO, and segmental SSDR, averaged over the four different noise types and six SNR conditions showing the baselines vs. the proposed *a priori* SNR estimators, and the **MMSE-LSA** spectral weighting rule.

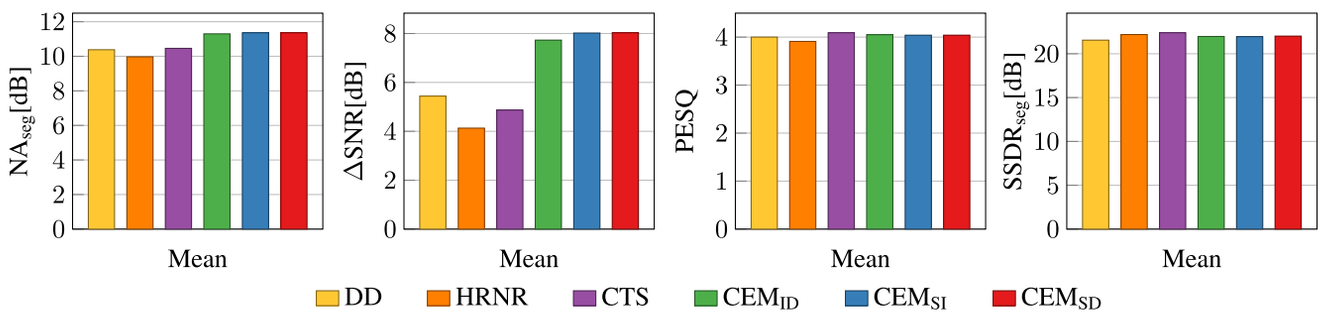


Fig. 12. Evaluation of segmental NA, ΔSNR, PESQ MOS-LQO, and segmental SSDR, averaged over the four different noise types and six SNR conditions showing the baselines vs. the proposed *a priori* SNR estimators, and the **SG-jMAP** spectral weighting rule.

other baselines, but SSDR_{seg} is significantly lower. The CEM approaches show highest ΔSNR in all conditions with comparable or better speech component quality w.r.t. the baselines. The flexion of the three curves also indicates a balanced working point in different SNR conditions. Using SG-jMAP (Fig. 8), the relationships across the tested algorithms are quite similar, except that HRNR improves in SSDR_{seg}, and gets worse in ΔSNR. The CEM algorithms obtain the best ΔSNR at very comparable speech component quality. For both spectral weighting rules, the CEM implementations among themselves show a very consistent rank order such that CEM_{ID} is marginally outperformed by the two template-based algorithms, which are almost equivalent.

Interestingly, all approaches obtain an increased preservation of the speech component quality with the SG-jMAP weighting rule compared to the MMSE-LSA spectral weighting rule at comparable (DD, CEM) or slightly lower (HRNR, CTS) ΔSNR values. *CEM is ahead of the baseline approaches by a ΔSNR of at least 2.35 dB (MMSE-LSA) and 2.29 dB (SG-jMAP) on total average.*

Figs. 9 and 10 provide the same analysis but for the SSDR_{seg} and the NA_{seg}. Fig. 9 depicting MMSE-LSA shows that CTS outperforms DD in terms of noise attenuation, however, with comparable quality of the speech component. The performance of HRNR is more difficult to interpret than before as the working point clearly is shifted since the single SNR conditions do not even roughly line up horizontally with the other baseline approaches under test. This results in a decreased speech component quality at substantially higher NA_{seg} values for each condition. Still, the proposed CEM approaches manage to show exceeding performance in all SNRs. The SG-jMAP spectral weighting rule is shown in Fig. 10. DD improves in NA_{seg} at similar SSDR_{seg} compared to Fig. 9 such that DD and CTS are much closer now. However, CTS is still able to consistently show a superior performance compared to the DD approach. Using SG-jMAP, HRNR loses performance in NA_{seg}, and becomes better in SSDR_{seg}, resulting in a clearer picture as opposed to Fig. 9. Best performing is again the CEM group, showing a similar behavior amongst themselves as in the other figures. In general, the range of NA_{seg} is the most compressed, considering the other two figures. *CEM is ahead of the baseline approaches by an NA_{seg} of up to 1.79 dB (MMSE-LSA) and 1.39 dB (SG-jMAP) on total average.*

We attribute the strong increase of NA_{seg} and ΔSNR in car noise mainly to the applied start decay as seen in Fig. 5, as it causes a good suppression in low frequencies typical for car noise. Moreover, for negative SNR conditions in pub noise we encounter cases where some other approach provides the best results with regard to the noise attenuation and speech component quality metrics. This is most likely due to F_0 estimation errors caused by the naïve pitch estimation which is unable to track a target speaker due to the presence of other speakers. We assume that the overall increase in NA_{seg} and also ΔSNR is caused by the introduced overestimation of the harmonics and the simultaneous attenuation in between them as shown in Fig. 4, center panel. Specifically, the attenuation between the harmonics should account for the increased overall noise attenuation. As there is usually a trade-off between noise attenuation and speech component quality [40], the CEM approach seems to mitigate this effect and allows us to be nearly on a par with the best baseline on average in terms of SSDR_{seg} and PESQ, while maintaining a higher NA_{seg} and ΔSNR.

https://doi.org/10.24355/dbbs.084-202102121510-0

To facilitate a conclusive interpretation of Tables I and II we provide bar charts depicting the overall mean values (last column of the mean section of Tables I and II for each measure) in Figs. 11 and 12 for the MMSE-LSA and the SG-jMAP weighting rule, respectively. Both figures show the advantage of the proposed approaches on average over the baselines in terms of $\mathrm{NA}_{\mathrm{seg}}$ and particularly ΔSNR. The CEM approaches are ahead of the baselines by at least 2 dB w.r.t. ΔSNR, while maintaining a very comparable speech component quality in both other measures, PESQ and $\mathrm{SSDR}_{\mathrm{seg}}$. However, the quality improvement from $\mathrm{CEM}_{\mathrm{ID}}$ to $\mathrm{CEM}_{\mathrm{SI}}$ or even $\mathrm{CEM}_{\mathrm{SD}}$ is only marginal. Nevertheless, $\mathrm{CEM}_{\mathrm{SD}}$ on average performs best among the CEM approaches. The HRNR approach seems to deliver a better speech component quality when used together with SG-jMAP (as compared to MMSE-LSA), which is strongly reflected in the $\mathrm{SSDR}_{\mathrm{seg}}$ measure at the cost of only a minor decrease in noise attenuation. Again, this shows how an advanced weighting rule can mend estimation flaws of earlier components in a noise reduction scheme. The DD and CTS baselines show a quite consistent performance regardless of the applied weighting rule.

In our experiments we encountered some empty templates during the training, which is caused by a lack of training material. A brief analysis has shown that mostly for lower cepstral bin indices we find every other set being empty, indicating that for some higher pitch frequencies ($F_0 > 400\,\mathrm{Hz}$) no material has been seen during training. However, we also obtain some coherent clusters for the lower frequencies. One way to avoid this could be to reduce the resolution of the cepstrum as we would not have seen any empty sets with the normal resolution but also would not have had the gain of the additional precision reflected by the coherent clusters. Furthermore, we could verify that an excitation template has been applied to 99.99% of the frames processed by the template-based methods, showing that empty templates do not have any significant relevance at this point. In addition to that, informal listening tests have shown that the proposed CEM methods also allow for almost musical tone-free noise suppression due to the instantaneous nature of *a priori* SNR estimation.[4]

VI. Conclusion

In this paper we have introduced three novel methods for instantaneous *a priori* SNR estimation utilizing the source-filter model for speech production. A preliminary denoised signal is decomposed into its source and corresponding filter, allowing to impose an idealized excitation on the degenerated source. The cepstral domain is exploited to manipulate the excitation signal at hand. We further enhance the technique by obtaining excitation templates from clean speech in either a speaker-independent or speaker-dependent fashion, where the latter is the slightly superior approach on average. However, the idealized technique shows some advantages over the codebook-based approaches, especially in SNR conditions ≥ 10 dB where it achieves equal or even better speech component quality at the cost of slightly lower noise suppression. We tested our algorithms and three baseline estimators in a common noise reduction algorithm with two different spectral weighting rules and managed to show a ΔSNR improvement of more than 2 dB, while the amount of speech distortion is largely kept on a constant level. Future work will include an enhancement of not only the excitation but also the envelope which is still taken from the preliminary denoised signal. Also a more sophisticated approach to F_0 estimation or tracking could improve our approaches in low-SNR conditions with multiple speakers.

[4]Audio samples can be found under: https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2017-taslp-cem/

Acknowledgment

The authors would like to thank T. Gerkmann for providing an executable to simulate the CTS baseline and also the anonymous reviewers for their constructive feedback helping to improve the quality of this paper.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[3] I. Cohen, "Speech enhancement using super-Gaussian speech models and noncausal *a priori* SNR estimation," *Speech Commun.*, vol. 47, no. 3, pp. 336–350, Nov. 2005.

[4] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[5] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel *a priori* SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Las Vegas, NV, USA, Mar. 2008, pp. 4897–4900.

[6] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to *a priori* SNR estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[7] S. Elshamy, N. Madhu, W. J. Tirry, and T. Fingscheidt, "An iterative speech model-based *a priori* SNR estimator," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 1740–1744.

[8] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[9] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved *a posteriori* speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[10] P. Scalart and J. V. Filho, "Speech enhancement based on *a priori* signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Atlanta, GA, USA, May 1996, pp. 629–632.

[11] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[12] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[13] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.

[14] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Montreal, QC, Canada, May 2004, pp. 289–292.

[15] M. Djendi and P. Scalart, "Reducing over- and under-estimation of the *a priori* SNR in speech enhancement techniques," *Digit. Signal Process.*, vol. 32, pp. 124–136, Sep. 2014.

[16] F. Deng and C. Bao, "Speech enhancement based on AR model parameters estimation," *Speech Commun.*, vol. 79, pp. 30–46, May 2016.

https://doi.org/10.24355/dbbs.084-202102121510-0

[17] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[18] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[19] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, no. 2, pp. 220–231, Feb. 2006.
[20] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[21] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.
[22] P. Vary, "Noise suppression by spectral magnitude estimation—Mechanism and theoretical limits—," *Signal Process.*, vol. 8, no. 4, pp. 387–400, Jul. 1985.
[23] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, Apr. 2011.
[24] P. Mowlaee, R. Saeidi, and Y. Stylianou, "INTERSPEECH 2014 special session: Phase importance in speech processing," in *Proc. INTERSPEECH*, Singapore, Sep. 2014, pp. 1623–1627.
[25] J. Kulmer and P. Mowlaee, "Phase estimation in single channel speech enhancement using phase decomposition," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 598–602, May 2015.
[26] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.
[27] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *Proc. 19th Eur. Signal Process. Conf.*, Barcelona, Spain, Sep. 2011, pp. 2309–2313.
[28] J. E. Markel and A. H. Gray, *Linear Prediction of Speech*. Berlin, Germany: Springer-Verlag, 1976.
[29] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin, Germany: Springer-Verlag, 2008.
[30] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
[31] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, Feb. 1967.
[32] "Super Wideband Stereo Speech Database," NTT Advanced Technology Corporation (NTT-AT). [Online]. Available: http://www.ntt-at.com/product/widebandspeech/
[33] European Telecommunications Standards Institute, *Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database*, ETSI EG 202 396-1, Sep. 2008.
[34] International Telecommunication Union, *Objective Measurement of Active Speech Level*, Telecommunication Standardization Sector (ITU-T), Rec. P.56, Dec. 2011.
[35] H. Yu, "Post-filter optimization for multichannel automotive speech enhancement," Ph.D. dissertation, Inst. Commun. Tech., Technische Univ. Braunschweig, Braunschweig, Germany, 2013.
[36] S. Gustafsson, R. Martin, and P. Vary, "On the optimization of speech enhancement systems using instrumental measures," in *Proc. Workshop Qual. Assessment Speech, Audio, Image Commun.*, Darmstadt, Germany, Mar. 1996, pp. 36–40.
[37] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008.
[38] International Telecommunication Union, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, Telecommunication Standardization Sector (ITU-T) Rec. P.862, Feb. 2001.
[39] International Telecommunication Union, *Narrow-Band Hands-Free Communication in Motor Vehicles*, Telecommunication Standardization Sector (ITU-T) Rec. P.1100, Jan. 2015.
[40] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

Samy Elshamy received the B.Sc. degree in bioinformatics from Friedrich-Schiller-Universität Jena, Jena, Germany, in 2011 and the M.Sc. degree in computer science from Technische Universität Braunschweig, Braunschweig, Germany, in 2013. He is currently working toward the Ph.D. degree in the field of speech enhancement at the Institute for Communications Technology, Technische Universität Braunschweig.

Nilesh Madhu received the Dr.-Ing. degree in electrical engineering and information technology from the Ruhr-Universität Bochum, Bochum, Germany, in 2009. Following this, he received a Marie-Curie fellowship for a two-year postdoctoral stay at the KU Leuven, Belgium, where he successfully applied his signal processing knowledge to the field of hearing prostheses. Since 2011, he has been with NXP, Leuven, Belgium, and is currently a Principal Scientist within the Product Line Mobile Audio Solutions, where he and his team work on developing innovative algorithms for audio and speech enhancement for mobile devices. He is passionate about signal processing, and is especially interested in the field of signal detection and enhancement for various applications, not just audio.

Wouter Tirry received the M.Sc. degree in physics and the Ph.D. degree in solar physics from the University of Leuven, Leuven, Belgium, in 1994 and 1998, respectively. As a Postdoc, he further pursued his research at the National Centre for Atmospheric Research, Boulder, CO, USA. Since 1999, he has been building up expertise in the domain of speech enhancement for mobile devices at Philips and NXP, Leuven, Belgium, as a Research Engineer and a System Architect. He is currently a Senior Principal at the Product Line Mobile Audio Solutions, NXP, leading the speech technology development activities.

Tim Fingscheidt (S'93–M'98–SM'04) received the Dipl.-Ing. degree in electrical engineering in 1993 and the Ph.D. degree in 1998 from RWTH Aachen University, Aachen, Germany. He further pursued his work on joint speech and channel coding as a Consultant in the Speech Processing Software and Technology Research Department, AT&T Labs, Florham Park, NJ, USA. In 1999, he entered the Signal Processing Department of Siemens AG (COM Mobile Devices) in Munich, Germany, and contributed to speech codec standardization in ETSI, 3GPP, and ITU-T. In 2005, he joined Siemens Corporate Technology in Munich, Germany, leading the speech technology development activities in recognition, synthesis, and speaker verification. Since 2006, he is a Full Professor in the Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany. His research interests include speech and audio signal processing, enhancement, transmission, recognition, and instrumental quality measures. He received several awards, among them a prize of the Vodafone Mobile Communications Foundation in 1999, and the 2002 prize of the Information Technology branch of the Association of German Electrical Engineers (VDE ITG), where he is leading the Speech Acoustics Committee ITG FA4.3 since 2015. From 2008 to 2010, he served as an Associate Editor for IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and since 2011 as a member of the IEEE Speech and Language Processing Technical Committee.

https://doi.org/10.24355/dbbs.084-202102121510-0

Publication III

S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "Two-Stage Speech Enhancement with Manipulation of the Cepstral Excitation," in *Proc. of HSCMA*, San Francisco, CA, USA, Mar. 2017, pp. 106–110

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

TWO-STAGE SPEECH ENHANCEMENT WITH MANIPULATION OF THE CEPSTRAL EXCITATION

Samy Elshamy, Nilesh Madhu°, Wouter Tirry° and Tim Fingscheidt**

*Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, D–38106 Braunschweig, Germany
°NXP Software, Interleuvenlaan 80, B–3001 Leuven, Belgium
{s.elshamy,t.fingscheidt}@tu-bs.de, {nilesh.madhu,wouter.tirry}@nxp.com

ABSTRACT

The development of new speech enhancement techniques is a continuous progress to combat the impairment of speech signals by various acoustical environmental influences. In this contribution we propose a new two-stage speech enhancement algorithm, exploiting the source-filter model to decompose a denoised target signal, and specifically we manipulate the excitation signal in the cepstral domain. The second stage therein is a refinement of the *a priori* signal-to-noise ratio (SNR) estimate used for the suppression gain calculation. Different to prior art, a higher noise attenuation can be achieved, without any more artifacts in the processed speech component.

Index Terms— *a priori* SNR, speech enhancement, cepstrum

1. INTRODUCTION

Speech enhancement has been and still is an important field of research to ensure proper speech quality even in adverse communication environments. Particularly single-channel speech enhancement with only a single noisy observation at hand faces several challenges. Common noise reduction schemes usually comprise a noise power estimate required for the *a priori* and the *a posteriori* SNR estimates, where one or both of the entities are usually controlling the calculation of suppression gains according to some spectral weighting rule.

The noise power estimate is provided by algorithms such as the minimum statistics (MS) approach [1], where the minimum value is tracked within a window of a smoothed input periodogram and the updating is not restricted to noise-only frames. Alternatively, the improved minima controlled recursive averaging (IMCRA) algorithm [2] could be used, where the tracked minimum of a smoothed input periodogram is indirectly controlling a smoothing parameter for the averaging of the input periodogram resulting in the required noise power estimate. The improved algorithm also tracks the minima during speech activity and introduces a compensation factor. A more recent and low complex method [3] utilizes a minimum mean-square error (MMSE) criterion to estimate the noise power and exhibits a low tracking delay making it especially attractive in non-stationary environments.

Several SNR estimation algorithms have been developed, amongst whose the well-known decision-directed (DD) approach by Ephraim and Malah [4] is found. It is a weighted sum with two components where the first is representing the SNR of the last frame's estimates for the clean speech and the noise, while the second is an *a priori* estimate obtained from the *a posteriori* SNR of the current frame. The weights in sum are constrained to unity.

A cepstral method has been proposed in [5] where the cepstrum of a speech power spectral density is smoothed with variable factors depending on the corresponding quefrency being able to distinguish easily between coefficients related to the envelope and the excitation. The approach can be further improved by applying an improved bias compensation after [6].

Along with the DD approach an MMSE short-time spectral amplitude estimator has been published [4] and extended to an MMSE log-spectral amplitude estimator (MMSE-LSA) [7], both being commonly used as spectral weighting rules. They are modeling the Fourier coefficients of the speech and noise process as statistically independent Gaussian random variables. Moving from a Gaussian to a super-Gaussian assumption and from MMSE to maximum a posteriori estimation, an improved spectral weighting rule has been proposed by Lotter and Vary, [8], briefly referred to as SG-jMAP.

Multi-stage speech enhancement algorithms have been developed in the past, ranging from two-stage [9, 10] to three-stage [11] approaches. The latter is based on [9], introducing an *a priori* SNR improvement based on a harmonic regeneration noise reduction (HRNR), realized by a non-linear function in the time domain to enhance the harmonic structure. The method is coupled with the Wiener filter as outlined in [12].

In this contribution we propose an *a priori* SNR refinement to recalculate a spectral weighting rule as a second stage on top of a common noise reduction scheme. The approach utilizes the *source-filter* model to separate the denoised signal from the first stage into its spectral envelope and excitation by linear predictive coding (LPC) analysis, followed by the transformation of the excitation signal to the cepstral domain. We apply a speaker-independent template-based cepstral excitation manipulation scheme to further enhance the already denoised excitation signal and show improvement over common noise reduction schemes. A more thorough presentation and analysis of the underlying cepstral manipulation technique is to be found in [13].

The paper is structured as follows: In Section 2 we introduce the new template-based cepstral processing methodology followed by the experimental evaluation in Section 3. We finally conclude our contribution in Section 4.

2. NEW TWO-STAGE SPEECH ENHANCEMENT

In this section we briefly present our notations followed by the new two-stage speech enhancement approach as depicted in Figure 1, serving as reference throughout the whole section.

2.1. First Stage Noise Reduction

We model the microphone signal $y(n)$ as a superposition of the clean speech signal $s(n)$ and the noise signal $d(n)$, both in the time do-

https://doi.org/10.24355/dbbs.084-202102121510-0

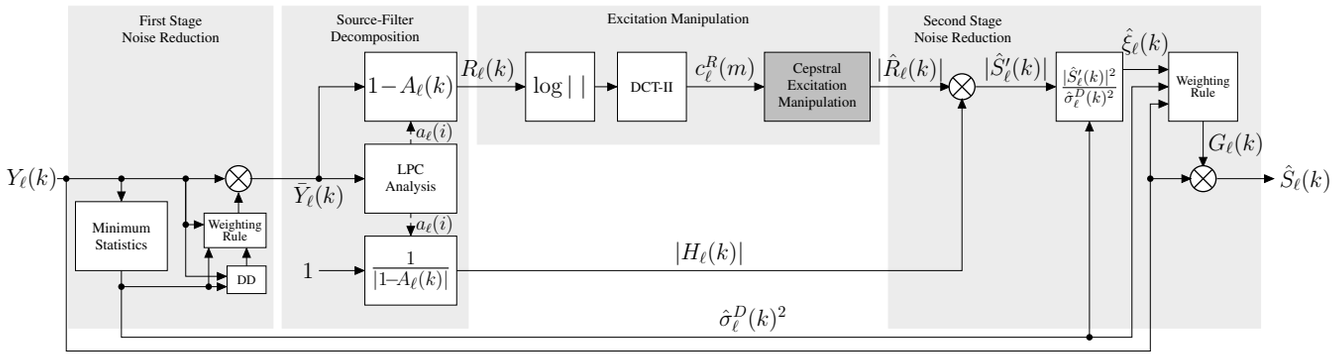


Fig. 1. Block diagram of the two-stage speech enhancement system. The first stage noise reduction consists of the minimum statistics (MS) noise power estimation algorithm, the decision-directed (DD) *a priori* SNR estimation approach, and a spectral weighting rule.

main with discrete-time sample index n, as $y(n) = s(n)+d(n)$. Applying the K-point discrete Fourier transform (DFT) we obtain the corresponding frequency domain representation $Y_\ell(k) = S_\ell(k) + D_\ell(k)$ with frame index ℓ and frequency bin index $0 \leq k \leq K-1$.

The first stage of the proposed technique (Figure 1, left light gray block) is depicting a common noise reduction scheme that yields a preliminary denoised signal $\bar{Y}_\ell(k) = Y_\ell(k) \cdot \bar{G}_\ell(k)$ where $\bar{G}_\ell(k)$ represents a real-valued gain function obtained by a spectral weighting rule. The bin-selective gain functions are usually depending on the *a priori* SNR $\bar{\xi}_\ell(k) = {\sigma_\ell^{\bar{S}}(k)^2}/{\sigma_\ell^D(k)^2}$, here estimated by the DD approach [4], and sometimes also on the *a posteriori* SNR $\gamma_\ell(k) = {|Y_\ell(k)|^2}/{\sigma_\ell^D(k)^2}$. The required noise power estimate $\hat{\sigma}_\ell^D(k)^2$ is calculated using the MS algorithm, where in general any suitable algorithm could be used.

2.2. Source-Filter Decomposition

Next, the preliminary denoised signal $\bar{Y}_\ell(k)$ is subject to LPC analysis (Figure 1, second light gray block) to retrieve the spectral excitation $R_\ell(k)$ and the spectral envelope $H_\ell(k)$. The required correlation coefficients are obtained by applying the inverse discrete Fourier transform (IDFT) as $\left(\varphi_\ell^{\bar{y},\bar{y}}(\lambda)\right)_{\lambda=0}^{K-1} = \text{IDFT}\{\left(|\bar{Y}_\ell(k)|^2\right)\}$. The first $N+1 < K$ elements $\varphi_\ell^{\bar{y},\bar{y}}(\lambda)$, $\lambda \in \{0,1,\ldots,N\}$ of $\left(\varphi_\ell^{\bar{y},\bar{y}}(\lambda)\right)_{\lambda=0}^{K-1}$ are used to calculate N LPC coefficients $a_\ell(i)$, $i \in \{1,2,\ldots,N\}$ by applying the Levinson-Durbin recursion. To obtain the LP analysis filter coefficients in the frequency domain ($1-A_\ell(k)$), the obtained N LPC coefficients are padded with $K-N-1$ zeros and subsequently transformed by a K-point DFT:

$$(A_\ell(k))_{k=0}^{K-1} = \text{DFT}\left\{(0, a_\ell(1), \ldots, a_\ell(N), 0, \ldots, 0)\right\}. \quad (1)$$

The excitation signal is then obtained as $R_\ell(k) = \bar{Y}_\ell(k) \cdot (1 - A_\ell(k))$ and correspondingly the inverse filter depicting the spectral envelope as $H_\ell(k) = \frac{1}{1-A_\ell(k)}$.

2.3. Cepstral Excitation Manipulation (CEM)

In order to further analyze the spectrum of the excitation signal, we apply the discrete cosine transform of type II (DCT-II) [14] to the logarithmic amplitude spectrum to obtain the cepstral coefficients as:

$$c_\ell^R(m) = \sum_{k=0}^{K-1} \log\left(|R_\ell(k)|\right) \cdot \cos\left[\pi m\,(k+0.5)\,\tfrac{1}{K}\right]. \quad (2)$$

The cepstral bin index is described by $m \in \mathcal{M} = \{0,1,\ldots,K-1\}$ and as we compute the cepstrum on the whole spectrum the resulting coefficients have a doubled resolution. After applying the cepstral excitation manipulation, the signal $c_\ell^{\hat{R}}(m)$ needs to be transformed back into the spectral amplitude domain by

$$|\hat{R}_\ell(k)| = \exp\left(\frac{c_\ell^{\hat{R}}(0)}{K} + \frac{2}{K}\sum_{m=1}^{K-1} c_\ell^{\hat{R}}(m) \cdot \cos\left[\pi m\,(k+0.5)\,\tfrac{1}{K}\right]\right). \quad (3)$$

Having manipulated the cepstrum of the excitation signal and after the IDCT-II, the second stage noise reduction starts by mixing it with the inherent envelope $H_\ell(k)$ as obtained after the first stage noise reduction to obtain an enhanced clean speech amplitude estimate $|\hat{S}'_\ell(k)| = |\hat{R}_\ell(k)| \cdot |H_\ell(k)|$. We finally use it to calculate a refined *instantaneous a priori* SNR as

$$\hat{\xi}_\ell(k) = \frac{|\hat{S}'_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}, \quad (4)$$

which is used in a second gain function together with the *a posteriori* SNR estimate from the first stage noise reduction. The clean speech estimate after the second stage noise reduction is obtained by applying the new real-valued gain function $G_\ell(k)$ to the *unfiltered microphone signal* $Y_\ell(k)$ as $\hat{S}_\ell(k) = Y_\ell(k) \cdot G_\ell(k)$.

The basic idea of our manipulation scheme is to replace the excitation after the first stage noise reduction by an excitation that has been trained beforehand on clean speech material according to the pitch and additionally boost the harmonic structure. Therefore, we exploit the convenience of the cepstral representation which allows to estimate the pitch bin index m_{F_0} by identifying the maximum amplitude [15] in a defined range of naturally occurring pitch frequencies $50\,\text{Hz} \leq F_0 \leq 500\,\text{Hz}$. Converting the frequencies to cepstral bin indices with $f = \frac{2f_s}{m}$ at a sampling rate $f_s = 8\,\text{kHz}$ yields the range $m \in \mathcal{M}_{F_0} = \{m_{500} = 32, \ldots, m_{50} = 320\}$. The desired bin index corresponding to the pitch frequency is then found as

$$m_{F_0} = \arg\max_{\mu \in \mathcal{M}_{F_0}} \left(c_\ell^R(\mu)\right). \quad (5)$$

We generate the speaker-independent excitation templates by analyzing some clean speech material and decomposing the clean speech spectrum $S_\ell(k)$ for each frame ℓ by LPC analysis into its spectral excitation and envelope (i.e., assuming $\bar{Y}_\ell(k) = S_\ell(k)$ in Figure 1 during training) as a first step. Next, the obtained spectral excitation is transformed into the cepstral domain by (2) and we define a set containing all the cepstral training vectors $\mathbf{c}_\ell^R$ for each possible pitch bin index $m \in \mathcal{M}_{F_0}$, identified by (5) as

$$\mathcal{C}_{m_{F_0}} = \left\{\mathbf{c}_\ell^R | c_\ell^R(m_{F_0}) \geq c_\ell^R(\mu)\ \forall \mu \in \mathcal{M}_{F_0}\right\}. \quad (6)$$

https://doi.org/10.24355/dbbs.084-202102121510-0

Consequently, we average each set $\mathcal{C}_{m_{F_0}}$ to obtain one representative cepstral excitation vector $\bar{\mathbf{c}}^R(m_{F_0})$ for each pitch bin index as

$$\bar{\mathbf{c}}^R(m_{F_0}) = \frac{1}{|\mathcal{C}_{m_{F_0}}|} \sum_{\mathbf{c}^R \in \mathcal{C}_{m_{F_0}}} \mathbf{c}^R, \tag{7}$$

where $|\cdot|$ depicts the cardinality of the set and the frame index ℓ has been dropped as it is no longer required.

These cepstral excitation vectors (templates) are utilized as follows. Having identified the pitch bin index of the current frame ℓ using (5), we define a new cepstral vector $c_\ell^{\hat{R}}(m)$, and in order to keep the same energy level as in the excitation signal after the first stage noise reduction, we transfer the cepstral energy coefficient ($m = 0$) into our new cepstral vector:

$$c_\ell^{\hat{R}}(0) = c_\ell^R(0). \tag{8}$$

Additionally, we overestimate the amplitude of the identified pitch bin index m_{F_0}, which boosts the harmonic structure and attenuates the valleys in between, and also use it in the new cepstrum:

$$c_\ell^{\hat{R}}(m_{F_0}) = c_\ell^R(m_{F_0}) \cdot \alpha_\ell(m_{F_0}) \tag{9}$$

where the overestimation factor $\alpha_\ell(m)$ may be chosen in a time-variant and bin-dependent fashion which has not been further examined in this paper. The remainder of the new cepstral vector is filled with amplitudes from the corresponding template (7) as

$$c_\ell^{\hat{R}}(m) = \bar{c}^R(m_{F_0}, m) \qquad \forall m \notin \{0, m_{F_0}\}. \tag{10}$$

3. EXPERIMENTAL EVALUATION

3.1. Experimental Setup

We conduct our experiments with a sample rate $f_s = 8\,\text{kHz}$, frame size of $K = 256$ samples, a frame shift of 50 % and a periodic square root Hann window is employed for both analysis and overlap-add synthesis. As clean speech database we utilize the NTT super wideband database [16] downsampled to 8 kHz. Only American and British English speakers are used (14 in total). The gender distribution is equal and we use 100 utterances per speaker, where 80 are assigned to a training set and the remaining 20 to a test set. The speaker-independent excitation template training is conducted in a leave-one-out fashion to compensate for the small amount of available training material by generating a template memory for each of the 14 speakers with the training data of the 13 other speakers. We use road, car, office, and pub noise from the ETSI database [17] and the segments used to generate noisy observations are randomly selected. We evaluate the algorithms in six different SNR conditions from -5 dB to 20 dB with an increment of 5 dB where the levels are adjusted as suggested in P.56 [18] amounting to a total of 6720 files per simulation. We evaluate the proposed two-stage speech enhancement technique with two different spectral weighting rules, either MMSE-LSA or SG-jMAP being used in *both* stages, against the corresponding output of the first stage noise reduction, and additionally the three-stage algorithm (HRNR) in [11]. Every gain function is lower-bounded to $G_{\min} = -15\,\text{dB}$ and the DD *a priori* SNR estimation is driven with optimal[1] parameters [19] for each of the weighting rules and additionally for the HRNR algorithm. The required overestimation factor in (9) has been empirically set to $\alpha_\ell(m_{F_0}) = 2$ and is time invariant.

[1] Optimal parameters for DD *a priori* SNR estimation with the two weighting rules, and the HRNR algorithm:

MMSE-LSA: $\beta_{\text{DD}} = 0.975, \xi_{\min} = -15\,\text{dB}$
SG-jMAP: $\beta_{\text{DD}} = 0.993, \xi_{\min} = -14\,\text{dB}$
HRNR: $\beta_{\text{DD}} = 0.985, \xi_{\min} = -15\,\text{dB}$.

3.2. Quality Measures

The basis of our quality measuring is the white-box approach proposed in [20]. With it and the linearity assumption we are able to not only process the microphone signal $Y_\ell(k)$ with an obtained gain function $G_\ell(k)$ but also the single components $S_\ell(k)$ and $D_\ell(k)$ which yields two new components after applying IDFT and overlap-add synthesis. We call them the *filtered* clean speech component $\tilde{s}(n)$ and the *filtered* noise component $\tilde{d}(n)$, respectively, where $\hat{s}(n) = \tilde{s}(n) + \tilde{d}(n)$.

We evaluate two measures for the noise component and two measures for the speech component. First, the segmental noise attenuation (NA) [21] representing an averaged local frame-wise ratio of the noise and the filtered noise component is obtained as

$$\text{NA}_{\text{seg}} = 10 \log_{10} \left[\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{NA}(\ell) \right], \tag{11}$$

with

$$\text{NA}(\ell) = \frac{\sum_{\nu=0}^{N-1} d(\nu + \ell N)^2}{\sum_{\nu=0}^{N-1} \tilde{d}(\nu + \ell N + \Delta)^2}.$$

Here, ℓ is representing a frame of size $N = 256$, the sample delay is compensated by Δ, and $\frac{1}{|\mathcal{L}|}$ is a normalization factor with $|\mathcal{L}|$ being the number of all frames. On a more global level we define

$$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}, \tag{12}$$

which basically depicts the difference between the output SNR measured on the filtered components $\tilde{s}(n)$ and $\tilde{d}(n)$ and the input SNR measured on the unfiltered components $s(n)$ and $d(n)$. An improved SNR is represented by a high ΔSNR. Speech component quality is measured by the segmental speech-to-speech-distortion ratio (SSDR) [21] calculated as

$$\text{SSDR}_{\text{seg}} = \frac{1}{|\mathcal{L}_1|} \sum_{\ell \in \mathcal{L}_1} \text{SSDR}(\ell) \tag{13}$$

where $\mathcal{L}_1$ is the set of speech active frames which are identified by an energy threshold-based voice activity detection. Values are limited to $[-10, 30]\,\text{dB}$ by

$$\text{SSDR}(\ell) = \max\left\{\min\left\{\text{SSDR}'(\ell), R_{\max}\right\}, R_{\min}\right\}$$

with

$$\text{SSDR}'(\ell) = 10 \log_{10} \left[\frac{\sum_{\nu=0}^{N-1} s(\nu + \ell N)^2}{\sum_{\nu=0}^{N-1} e(\nu + \ell N)^2} \right]$$

and speech distortion

$$e(\nu + \ell N) = \tilde{s}(\nu + \ell N + \Delta) - s(\nu + \ell N).$$

Good speech component quality is reflected in a high segmental SSDR. As a second measure we use the PESQ mean opinion score (MOS-LQO) [22] which is applied to the *filtered* clean speech component $\tilde{s}(n)$ with the clean speech component as a reference. We do *not* measure PESQ on the enhanced signal $\hat{s}(n)$ since PESQ has not been validated for artifacts caused by noise reduction techniques. In line with P.1100 [23, Sect. 8] and using [20] to obtain the processed clean speech component, we instead measure the distortion of the clean speech component, thereby being also compliant to the intended use case of P.862 [22]. We simplify the evaluation of the four measures by introducing a figure of merit

$$\text{FoM} = \frac{1}{4} \frac{\text{NA}_{\text{seg}}}{\overline{\text{NA}_{\text{seg}}}} + \frac{1}{4} \frac{\Delta\text{SNR}}{\overline{\Delta\text{SNR}}} + \frac{1}{4} \frac{\text{PESQ}}{\overline{\text{PESQ}}} + \frac{1}{4} \frac{\text{SSDR}_{\text{seg}}}{\overline{\text{SSDR}_{\text{seg}}}}, \tag{14}$$

https://doi.org/10.24355/dbbs.084-202102121510-0

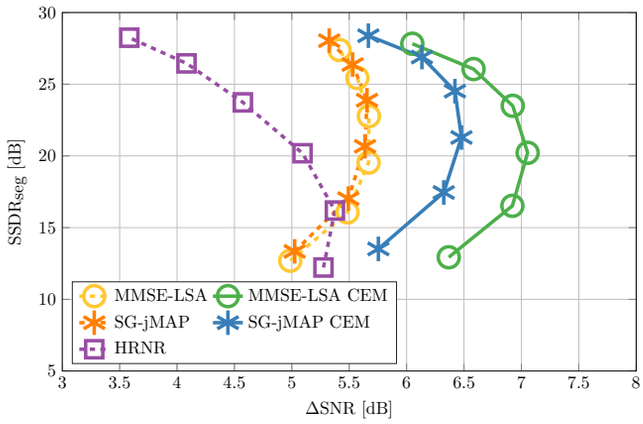


Fig. 2. Segmental SSDR and **ΔSNR** averaged over four different noise types for the different algorithms under test.

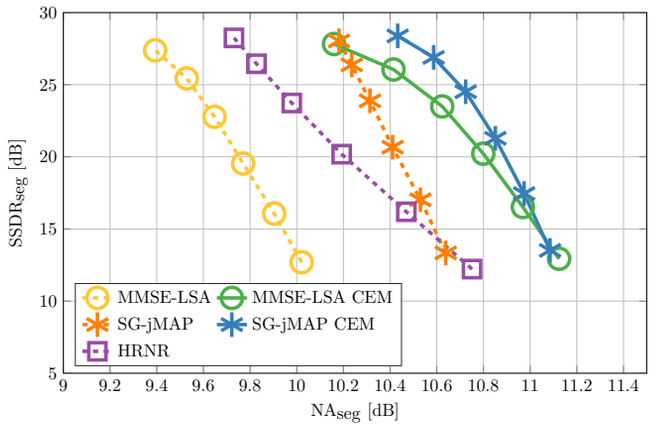


Fig. 3. Segmental SSDR and **segmental NA** averaged over four different noise types for the different algorithms under test.

which incorporates each measure equally and allows to draw conclusions about the balancedness of the compared approaches. Here, the normalizing component is the average of the respective measure over the approaches for all SNRs and noise types separately. A high value indicates a good balance of the four measures.

3.3. Experimental Evaluation

In Figure 2 we plot the segmental SSDR over the ΔSNR for each of the tested algorithms averaged over the four noise types. Each marker depicts one SNR condition from -5 dB to 20 dB in steps of 5 dB (bottom to top). It is clearly visible that the proposed two-stage approaches (solid lines ∘ and ∗) outperform their corresponding single-stage algorithms (dashed lines ∘ and ∗) by obtaining up to more than 1 dB higher ΔSNR, simultaneously achieving a comparable speech component quality. The HRNR approach performs a bit better than the single-stage approaches in low-SNR conditions, but is in all other conditions still clearly inferior to the proposed approaches. Figure 3 depicts the segmental SSDR over the NA$_{seg}$ and confirms the advantage of the proposed two-stage algorithms over both the single-stage and the HRNR approaches. It is worth to note that among the reference approaches SG-jMAP seems to provide the best overall trade-off between quality of the speech component and noise attenuation. The CEM algorithms clearly show superior performance over their corresponding single-stage systems in terms of NA$_{seg}$, while again obtaining comparable speech component quality. Table 1 shows the FoM for each condition separately including a mean over the SNR conditions and also the noise types. This allows to consider the four measures at once and to deduce information on the balancedness of the approaches. The two-stage CEM approaches clearly obtain the best scores in all conditions. They particularly show a significant improvement in car noise, office noise, and to some extent also in pub noise. With respect to the average values, both CEM approaches perform almost equally well which shows that the proposed methods clearly operate with a better final *a priori* SNR estimate compared to the three reference algorithms: This reduces the importance of choosing one or the other second stage spectral weighting rule. Informal listening tests[2] have shown that an explicit discrimination between voiced or unvoiced frames for the application of the algorithm is not vital.

[2]Audio samples can be found under:
https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2017-hscma/

4. CONCLUSION

In this paper we have introduced a new two-stage speech enhancement technique utilizing the source-filter method to decompose a preliminary denoised signal. We manipulate the excitation signal in the cepstral domain to obtain an enhanced clean speech estimate that is used for a refined *a priori* SNR estimation. We tested the CEM method with two different spectral weighting rules and could show consistent improvement over corresponding single-stage approaches and a further multi-stage reference not only by means of a condensed FoM but also in some typical quality measures.

Table 1. Evaluating the FoM for four different noise types, six SNR conditions, the references vs. the two-stage CEM approaches.

		FoM						
	SNR [dB]	-5	0	5	10	15	20	mean
ROAD	MMSE-LSA	0.86	0.92	0.97	1.02	1.05	0.98	0.97
	SG-jMAP	0.90	0.95	1.01	1.06	1.09	1.02	1.01
	HRNR	0.88	0.93	0.98	1.02	1.05	0.98	0.97
	MMSE-LSA CEM	0.92	0.98	1.03	1.06	1.08	1.03	1.02
	SG-jMAP CEM	**0.93**	**0.99**	**1.05**	**1.09**	**1.11**	**1.04**	**1.03**
CAR	MMSE-LSA	0.90	0.95	0.97	0.99	1.00	0.97	0.96
	SG-jMAP	0.92	0.96	0.99	1.01	1.01	0.98	0.98
	HRNR	0.92	0.94	0.93	0.91	0.89	0.91	0.92
	MMSE-LSA CEM	**1.04**	**1.08**	**1.10**	**1.11**	**1.10**	**1.09**	**1.09**
	SG-jMAP CEM	1.00	1.04	1.07	1.08	1.07	1.05	1.05
OFFICE	MMSE-LSA	0.78	0.89	0.98	1.04	1.09	0.98	0.96
	SG-jMAP	0.80	0.91	0.99	1.06	1.10	1.00	0.98
	HRNR	0.74	0.84	0.92	0.97	1.01	0.92	0.90
	MMSE-LSA CEM	**0.92**	**1.05**	**1.14**	**1.20**	**1.23**	**1.13**	**1.11**
	SG-jMAP CEM	0.86	0.99	1.08	1.14	1.17	1.07	1.05
PUB	MMSE-LSA	0.74	0.89	1.01	1.09	1.16	1.02	0.98
	SG-jMAP	0.74	0.90	1.02	1.11	1.18	1.03	0.99
	HRNR	0.67	0.83	0.94	1.03	1.10	0.95	0.92
	MMSE-LSA CEM	**0.75**	**0.97**	**1.10**	**1.19**	**1.25**	**1.09**	**1.06**
	SG-jMAP CEM	**0.75**	0.95	1.08	1.17	1.24	1.08	1.05
Means	MMSE-LSA	0.82	0.91	0.98	1.04	1.07	0.99	0.97
	SG-jMAP	0.84	0.93	1.00	1.06	1.10	1.01	0.99
	HRNR	0.80	0.88	0.94	0.98	1.01	0.94	0.93
	MMSE-LSA CEM	**0.91**	**1.02**	**1.09**	**1.14**	**1.17**	**1.08**	**1.07**
	SG-jMAP CEM	0.88	0.99	1.07	1.12	1.15	1.06	1.05

https://doi.org/10.24355/dbbs.084-202102121510-0

5. REFERENCES

[1] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[2] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[3] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[5] C. Breithaupt, T. Gerkmann, and R. Martin, "A Novel A Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing," in *Proc. of ICASSP*, Las Vegas, NV, USA, Mar. 2008, pp. 4897–4900.

[6] T. Gerkmann and R. Martin, "On the Statistics of Spectral Amplitudes After Variance Reduction by Temporal Cepstrum Smoothing and Cepstral Nulling," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[7] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[8] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[9] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A Two-Step Noise Reduction Technique," in *Proc. of ICASSP*, Montreal, Quebec, Canada, May 2004, pp. 289–292.

[10] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy Speech Enhancement Using Harmonic-Noise Model and Codebook-Based Post-Processing," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1194–1203, May 2007.

[11] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[12] P. Scalart and J. V. Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP*, Atlanta, GA, USA, May 1996, pp. 629–632.

[13] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous A Priori SNR Estimation by Cepstral Excitation Manipulation," *Submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[14] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2001.

[15] A. M. Noll, "Cepstrum Pitch Determination," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, Feb. 1967.

[16] "Super Wideband Stereo Speech Database," NTT Advanced Technology Corporation (NTT-AT).

[17] ETSI, *EG 202 396-1: Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database*, European Telecommunications Standards Institute, Sept. 2008.

[18] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.

[19] H. Yu, *Post-Filter Optimization for Multichannel Automotive Speech Enhancement*, Ph.D. thesis, Technische Universität Braunschweig, 2013.

[20] S. Gustafsson, R. Martin, and P. Vary, "On the Optimization of Speech Enhancement Systems Using Instrumental Measures," in *Proc. of Workshop on Quality Assessment in Speech, Audio, and Image Communication*, Darmstadt, Germany, Mar. 1996, pp. 36–40.

[21] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-Optimized Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, May 2008.

[22] ITU, *Rec. P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Feb. 2001.

[23] ITU, *Rec. P.1100: Narrow-Band Hands-Free Communication in Motor Vehicles*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Jan. 2015.

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

Publication IV

S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "DNN-Supported Speech Enhancement With Cepstral Estimation of Both Excitation and Envelope," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2460–2474, Dec. 2018

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

DNN-Supported Speech Enhancement With Cepstral Estimation of Both Excitation and Envelope

Samy Elshamy, Nilesh Madhu, Wouter Tirry, and Tim Fingscheidt, *Senior Member, IEEE*

***Abstract*—In this paper, we propose and compare various techniques for the estimation of clean spectral envelopes in noisy conditions. The source-filter model of human speech production is employed in combination with a hidden Markov model and/or a deep neural network approach to estimate clean envelope-representing coefficients in the cepstral domain. The cepstral estimators for speech spectral *envelope*-based noise reduction are both evaluated alone and also in combination with the recently introduced cepstral *excitation* manipulation (CEM) technique for *a priori* SNR estimation in a noise reduction framework. Relative to the classical MMSE short time spectral amplitude estimator, we obtain more than 2 dB higher noise attenuation, and relative to our recent CEM technique still 0.5 dB more, in both cases maintaining the quality of the speech component and obtaining considerable SNR improvement.**

***Index Terms*—*a priori* SNR, speech enhancement.**

I. Introduction

SPEECH enhancement is an important field of research to aid the most natural way of communication for human beings. It comprises a variety of applications among them dereverberation, acoustic echo cancellation, artificial bandwidth extension, voice activity detection, speech presence probability estimation, and also noise reduction algorithms. Many of these applications require the estimation of an *a priori* SNR which we are investigating in this publication in the context of a noise reduction framework. Furthermore, we focus on approaches exploiting the cepstral domain, since its properties and advantages have gained considerable attention in the recent past. For each component of a common noise reduction scheme, such as noise power estimator, *a priori* SNR estimator, and spectral weighting rule, approaches have been developed that exploit the cepstral domain.

Manuscript received April 10, 2018; revised July 11, 2018 and August 21, 2018; accepted August 22, 2018. Date of publication August 31, 2018; date of current version September 14, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andy W. H. Khong. *(Corresponding author: Tim Fingscheidt.)*

S. Elshamy and T. Fingscheidt are with the Institute for Communications Technology, Technische Universität Braunschweig, 38106 Braunschweig, Germany (e-mail: s.elshamy@tu-bs.de; t.fingscheidt@tu-bs.de).

N. Madhu was with NXP Software, 3001 Leuven, Belgium. He is now with the Internet Technology and Data Science Lab, Universiteit Gent-imec, 9052 Gent, Belgium (e-mail: nilesh.madhu@ugent.be).

W. Tirry is with NXP Software, 3001 Leuven, Belgium (e-mail: wouter.tirry@nxp.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2018.2867947

A *noise power estimation* algorithm based on minimum mean-square error (MMSE) estimation originally proposed in [1] has been further improved by Gerkmann *et al.* in [2] by a bias compensation which is required due to the necessity of estimating intermediate entities and therewith arising aberrations. Finally, the estimator has been augmented with temporal cepstrum smoothing [3] to enhance the speech power estimation, resulting in higher noise attenuation and thus improving the signal-to-noise ratio (SNR) [4].

A cepstral *a priori SNR estimation* technique has been proposed in [5], where the ability to easily address the fine structure and the envelope of a speech power spectral density separately, has been successfully exploited by applying different smoothing factors to the corresponding regions in the cepstral domain. Here, also an improved bias compensation [6] can be employed to further increase the performance.

We presented a cepstral excitation manipulation (CEM) method in [7], [8] that benefits from the direct modeling of the excitation signal. It is obtained via linear predictive coding (LPC) analysis and is subsequently replaced by a pitch-dependent excitation template which has been extracted from clean speech prior to its application. The approach successfully conquers an often reported issue with low-order models considering the shortcomings of noise suppression between the spectral harmonics [9], [10]. Furthermore, it renders means such as a voicing-sensitive postfilter, spectral mask, or speech presence probability estimation [9], [11], needless.

The last component of common noise reduction schemes, a *spectral weighting rule*, has been published by Breithaupt *et al.* in [3], performing smoothing in the cepstral domain to finally suppress the noise in a noisy signal. It allows to successfully suppress spectral outliers that otherwise would cause musical tones. It is to say that the general concept of temporal cepstrum smoothing has found various applications in speech enhancement.

The source-filter model of human speech production, separating a speech signal into its excitation and envelope has also found its applications in speech enhancement and is used in various degrees. The usage of speech and noise spectral shapes as a priori information for speech enhancement has been suggested by Srinivasan *et al.* in [12]–[14] and was developed further over time. Two low-rank codebooks trained on speech and noise spectral shapes are employed and a maximum-likelihood (ML) estimate of the corresponding parameters, two indices for the codebook entries and two corresponding gain factors, given the noisy observation, are calculated. The obtained parameters are

https://doi.org/10.24355/dbbs.084-202102121510-0

used to estimate the spectra of speech and noise, and are finally used in a Wiener filter to calculate spectral weighting gains. A continuation of this work has been published by Rosenkranz *et al.* where cepstral modeling is preferred over autoregressive (AR) modeling [15].

A non-negative matrix factorization approach representing a source-filter system where separate dictionaries for the excitation and the envelopes are trained is proposed in [16]. During test it also requires a preliminary denoised signal as the algorithm needs additional information from the signal such as a pitch estimate. It seems to be quite complex and it is not entirely clear, whether it is a realtime-capable algorithm or not, at least the used pitch estimator [17] indicates that it is not suitable for telephony applications.

The approaches exploiting cepstro-temporal smoothing [3] address the source-filter model in a fashion that the cepstral coefficients are assigned to either part of the model, depending on their position in the cepstrum, and are treated differently. Please note that this kind of model is not subject to specific constraints as in LPC analysis, where a given order strictly defines the number of poles in the z-transform of the model.

A hidden Markov model (HMM) has been used for speech enhancement in [18], [19]. Therein, two HMMs are utilized to model the clean speech and the noise signal separately by AR processes. In both references, a Wiener filter is derived by incorporating the estimated spectral prototypes provided by the HMMs. Different from [18], [19] decouples the gain factors from the prototypes and introduces an explicit modeling of the gains, leading to a consistent improvement. The low-order modeling of the speech HMM suggests that the approaches also suffer from the same incapability to model the fine structure appropriately and thus leaves room for improvement.

With deep learning strategies on the rise, deep neural networks (DNNs) also find their way into speech enhancement and allow for a very broad variety of applications. Approaches range from directly estimating clean time-domain signals from the noisy observation [20] to mapping functions for extracted noisy features to clean features [21]. Those DNN techniques have in common that they completely disregard statistical speech enhancement approaches, which still are commonly utilized, and instead highly depend on their training material. However, it is also possible to incorporate DNNs into well-known statistical frameworks as, e.g., it has been shown in [22] that incorporating DNNs into a common noise reduction scheme and replacing certain estimators of the system yields better results than employing a simple regression DNN to estimate the clean speech directly. Source-filter model approaches for artificial speech bandwidth extension have been very successfully shown to take profit from DNN envelope modeling with or even without HMM [23]–[25]. Also, as known from automatic speech recognition, Gaussian mixture models (GMMs) have been successfully replaced by DNNs for the acoustic modeling [26].

In this publication, we investigate various approaches for the estimation of clean spectral envelopes based on noisy observations. In all cases, the actual estimation domain is the real-valued cepstrum, since it advantageously allows the minimum mean squared error (MMSE) as cost function. We evaluate the performance with respect to their application in *a priori* SNR estimation for a noise reduction framework, as we expect quite some benefit from envelope enhancement in this field. We start with utilizing a classical HMM driven by GMMs, which are subsequently replaced by a DNN. Furthermore, we also investigate the replacement of the entire HMM by a single DNN, which is providing posterior probabilities instead of likelihoods for the HMM, or the use of a DNN to estimate clean coefficients directly in regression mode. Finally, we combine the enhanced spectral envelope with our recently proposed CEM approach and incorporate it into the *a priori* SNR estimator from [8]. Note, however, that the field of application for the proposed spectral envelope estimators is not limited to these specific use cases.

In the following, we briefly introduce the signal model in Section II and revisit the cepstral *excitation* manipulation technique in Section III, followed by the investigation of our various methods for clean spectral envelope estimation based on a preliminary denoised signal in Section IV, where we gradually replace the HMM by a DNN. We describe our experimental setup in Section V-A and subsequently provide our simulations and evaluation in Section VI. We finally conclude this article in Section VII.

II. Signal Model

To model the microphone signal $y(n)$ we assume that the speech signal $s(n)$ and the noise signal $d(n)$ are superimposed in the time domain as

$$y(n) = s(n) + d(n), \tag{1}$$

where n is the discrete-time sample index. A corresponding frequency domain representation after a K-point discrete Fourier transform (DFT) is obtained as

$$Y_\ell(k) = S_\ell(k) + D_\ell(k), \tag{2}$$

with frame index ℓ and frequency bin index $0 \leq k \leq K-1$. Also, we assume statistical independence of the speech and noise signal, and that they have zero mean.

III. Cepstral Excitation Manipulation Baseline

We choose to utilize our recently published *a priori* SNR estimation and noise reduction framework [8], as its modularity allows us to easily integrate the proposed estimators and evaluate their performance either alone (dubbed "solo") or in interaction with the CEM approach (called "duo").

As depicted in Fig. 1, a preliminary noise reduction stage is employed to get a more suitable signal for the proposed methods. This first noise reduction stage is a common noise reduction scheme with a noise power estimator such as minimum statistics (MS) [27], improved minima controlled averaging [28], or a more recent approach, the unbiased MMSE-based estimator [2]. Subsequently, it is followed by an *a priori* SNR estimator, e.g., the decision-directed (DD) approach [29], and finally, a spectral weighting rule to calculate the real-valued gain factors in the frequency domain for the noise suppression. Some quite often used spectral weighting rules are, e.g., the

https://doi.org/10.24355/dbbs.084-202102121510-0

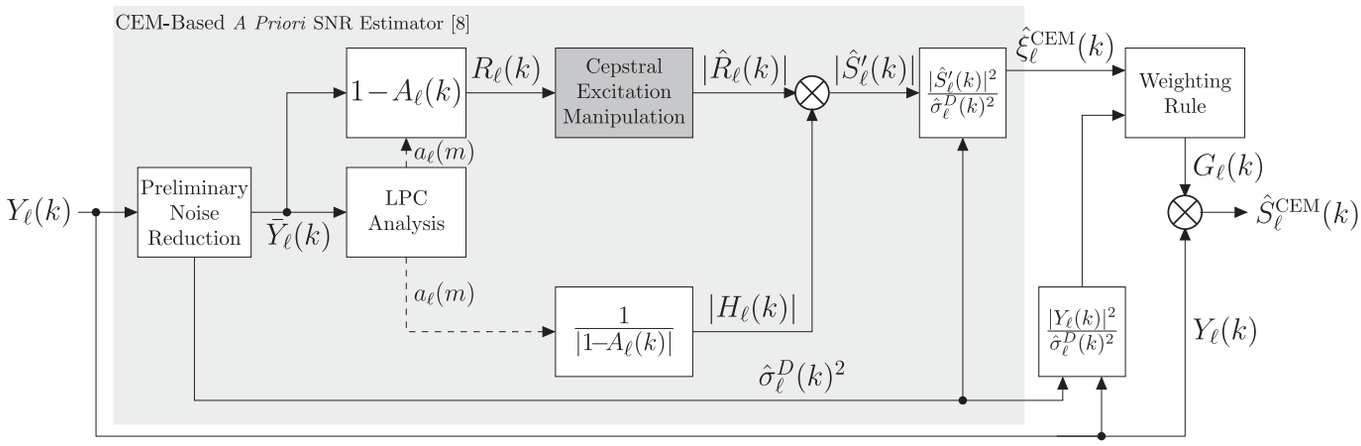


Fig. 1. High-level block diagram of the **cepstral excitation manipulation (CEM) noise reduction baseline**, incoporating a CEM-based *a priori* SNR estimator [8]. The proposed cepstral estimators for the spectral envelope are later on embedded into this approach (see Figs. 2 and 3).

MMSE log-spectral amplitude (MMSE-LSA) estimator [29], a more advanced gain function under a super-Gaussian assumption namely the super-Gaussian joint maximum a posteriori amplitude estimator [30], [31], or a simple Wiener filter [32]. In general, we do not restrict ourselves to a specific configuration, but have found a setup using MS noise power estimation along with the DD *a priori* SNR estimator and the MMSE-LSA spectral weighting rule as suitable for our method.

The preliminary denoised signal $\bar{Y}_\ell(k)$ is subsequently subject to a source-filter decomposition block where LPC analysis is utilized to obtain an excitation signal $R_\ell(k)$ and the corresponding envelope $H_\ell(k)$, separately. They relate to the preliminary denoised signal as

$$\bar{Y}_\ell(k) = R_\ell(k) \cdot H_\ell(k). \tag{3}$$

The CEM baseline as presented in [8] deals only with the enhancement of the excitation signal (Fig. 1, LPC analysis, upper path). As a first step of the CEM algorithm, the log-spectrum of the excitation signal is transformed into the cepstral domain by a discrete cosine transform of type II (DCT-II). Next, a (surprisingly) robust pitch estimation algorithm based on [33] provides the system with the corresponding cepstral pitch bin m_{F_0} by picking the maximum cepstral value within a quefrency bin range representing typical pitch frequencies.

Consequently, a pitch bin-dependent cepstral excitation template $c_\ell^{\hat{R}}(m)$, with m being the cepstral bin index, is selected from a template codebook that has been trained on clean speech residual signals. The designated template vector is subject to two major manipulations: First, the template's cepstral energy coefficient $c_\ell^{\hat{R}}(0)$ is replaced by the corresponding value of the preliminary denoised signal's residual $c_\ell^R(0)$ as

$$c_\ell^{\hat{R}}(0) = c_\ell^R(0) \tag{4}$$

in order to receive a signal with a similar power level as the input signal. Second, the cepstral amplitude of the pitch bin $c_\ell^R(m_{F_0})$ is also transferred into the already power-adjusted excitation template and subsequently overestimated by a factor $\alpha > 1$ as

$$c_\ell^{\hat{R}}(m_{F_0}) = \alpha \cdot c_\ell^R(m_{F_0}). \tag{5}$$

Thereby, the harmonic structure of the excitation signal is overemphasized in both directions: The positive and also the negative half waves experience a boost or an attenuation, respectively. As a result, the algorithm is able to retain weak harmonics which might have been corrupted by the preliminary denoising stage and additionally, achieve a higher noise attenuation between the harmonics. Both characteristics are depicting the core features of the CEM algorithm. Until now, the manipulated template is transformed back into the spectral domain by an inverse DCT-II and used further with the spectral amplitudes of the preliminary denoised signal's envelope $|H_\ell(k)|$ to provide an improved clean speech amplitude estimate $|\hat{S}'_\ell(k)|$ by mixing the two components as

$$|\hat{S}'_\ell(k)| = |\hat{R}_\ell(k)| \cdot |H_\ell(k)|. \tag{6}$$

Finally, it is used as the numerator for a refined *a priori* SNR estimate along with the obtained noise power estimate from the preliminary noise reduction

$$\hat{\xi}_\ell^{\mathrm{CEM}}(k) = \frac{|\hat{S}'_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}. \tag{7}$$

For more details about the CEM-based *a priori* SNR estimator, the interested reader may consult [8]. In *this* work, this estimator is embedded into a noise reduction framework as shown in Fig. 1, comprising also the computation of an *a posteriori* SNR

$$\gamma_\ell(k) = \frac{|Y_\ell(k)|^2}{\sigma_\ell^D(k)^2}, \tag{8}$$

as many gain functions $G_\ell(k)$ require either or both of the two SNRs for their calculation as

$$G_\ell(k) = f\left(\xi_\ell(k), \gamma_\ell(k)\right). \tag{9}$$

IV. Cepstral Estimation of the Envelope

In this section we will now present our new methods of cepstral estimation to obtain speech spectral envelopes under noisy conditions. As outlined in Section I, we will embed these estimators into a noise reduction baseline which already performs cepstral estimation of the speech residual (see Fig. 1). Note that

https://doi.org/10.24355/dbbs.084-202102121510-0

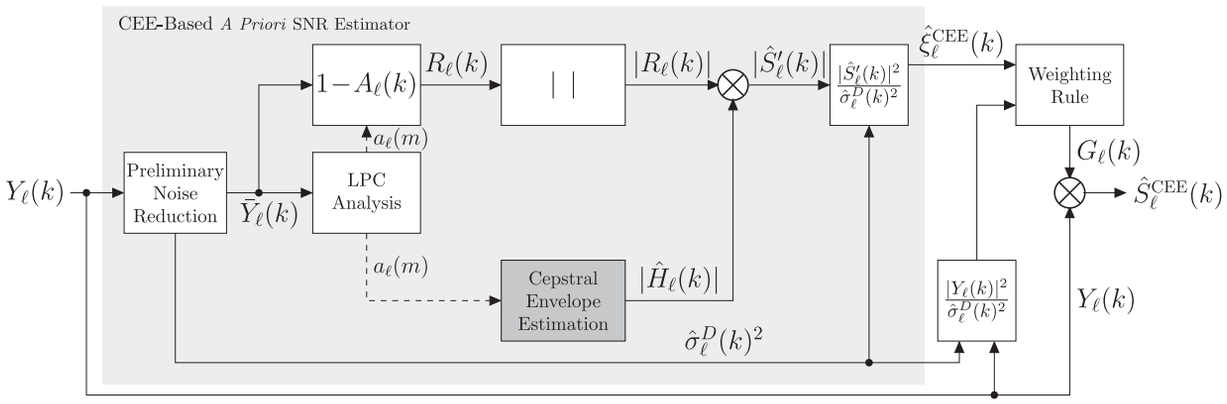


Fig. 2. High-level block diagram of the **proposed cepstral envelope estimation (CEE) noise reduction**, incorporating the new CEE-based *a priori* SNR estimator. For details of the CEE block please refer to Fig. 3.

this is only *one* of the many possibilities to employ our cepstral estimators of the speech spectral envelope. Our general approach advantageously uses a preliminary noise reduction, which provides an improved SNR for the subsequent envelope estimation. The spectral envelope of the preliminary denoised signal still suffers from distortions which tend to impede the speech quality, thus leaving room for further improvement. To our understanding it is reasonable to break down the noise reduction task for speech enhancement into smaller parts where possible. This is in line with divide-and-conquer strategies which have resulted in many useful solutions for various problems. As the production of speech can be modeled by two components, i.e., the source and the filter, it appears logical to attend each at a time which also has been done in, e.g., [13].

As a general framework we decided to employ a hidden Markov model (HMM) in order to estimate a clean spectral envelope, given the preliminary denoised observation. The motivation behind this is that we want to move from a bin-individual *a priori* SNR estimation (e.g., as the DD approach provides) to a more coherent and inter-frequency-dependent solution. Given the limited DFT length, this should be closer to the actual relationship between frequencies in speech, since they are not completely independent [34]. When dealing with spectral envelopes, this inter-frequency dependence becomes even more obvious. The application of a codebook that has been trained on clean speech spectral envelopes should be able to provide envelopes with a more realistic dependency between the *frequency* bins. In addition to that, we expect the HMM to capture the *temporal* context of envelopes which are usually smooth in transition.

The HMM in its classic form is using Gaussian mixture models (GMMs) to model the emission probabilities. As a second approach and along with the trend of deep learning we also employ a deep neural network (DNN) for classification to replace the GMMs. It has been shown in [26] that DNNs are capable of providing higher classification rates than GMMs, especially for acoustic models. A third variant we propose omits the HMM completely and solely uses the posterior distribution delivered by the classification DNN. As a fourth option we present a regression DNN in order to directly estimate clean envelope coefficients from the preliminary denoised observation.

In the following, we provide a generic recipe for our framework and the required training processes, while distinct parameters of our experimental setup will be provided in Section V-A.

A. Feature Conversion

As can be seen in Fig. 2, we also operate on the preliminary denoised signal $\bar{Y}_\ell(k)$, which is decomposed into its source and filter by means of an LPC analysis. Up to this stage in the block diagram, both approaches CEM and also the now introduced cepstral envelope estimation (CEE) share the same processing structure. Now, the difference is that we operate on the LPC coefficients modeling the envelope (Fig. 2, LPC analysis, lower path) and not the excitation signal as before (Fig. 1, LPC analysis, upper path).

Since some training processes require the averaging of feature vectors, using the LPC coefficients directly could lead to instabilities. To obtain a representation of the envelope that has more suitable mathematical properties (Fig. 3, feature conversion block), we convert the N LPC coefficients to $N+1$ cepstral coefficients by the following two formulae [35], which have been adjusted to our notation[1] of the LP analysis filter, here and also in (17), as (superscript *H* stands for the envelope filter)

$$c_\ell^H(m=0) = 0 = \log(P_p = 1). \tag{10}$$

The prediction error power P_p is set to an arbitrary fixed value to have envelopes with equal energy, allowing us to reduce the feature dimension to N since the zeroth coefficient is always zero. For $1 \le m \le N$ we calculate the cepstral coefficients recursively by

$$c_\ell^H(m) = a_\ell(m) + \frac{1}{m}\sum_{\mu=1}^{m-1}\left[(m-\mu)\cdot a_\ell(\mu)\cdot c_\ell^H(m-\mu)\right]. \tag{11}$$

We only compute the $N+1$ non-redundant cepstral coefficients to maintain a small dimension, omitting $c_\ell^H(m=0)$ as explained and thus work with N features.

[1] We denote the LPC analysis filter as $H_\ell(k) = 1 - A_\ell(k)$.

https://doi.org/10.24355/dbbs.084-202102121510-0

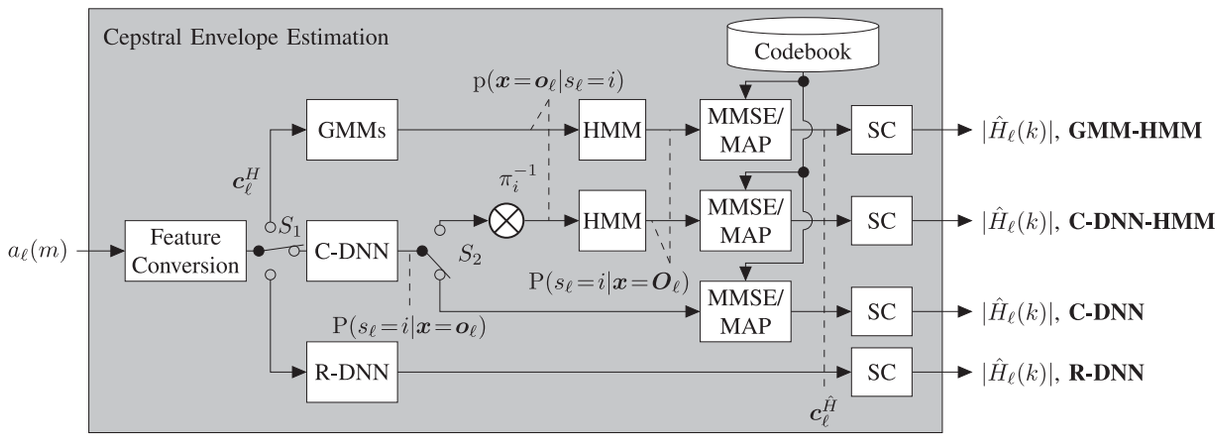


Fig. 3. Block diagram of the four different **proposed approaches for cepstral envelope estimation (CEE)** using either an HMM in combination with GMMs (first path) or alternatively a classification DNN to model the emissions (second path), or to model the posterior distribution of a classification DNN directly (third path). All these three approaches work with an LBG codebook for clean cepstral envelopes. Another option is a DNN trained as regressor (fourth path), estimating the clean cepstral coefficients directly from the input features. Since each approach yields enhanced cepstral coefficients, a required conversion to the spectral domain takes place in the spectral conversion (SC) boxes. Any of the four methods shown on the right side of the figure is determined by the setting of the switches S_1 and S_2 (shown: **C-DNN**).

In order to remove channel mismatches, we normalize all data in a bin-wise manner by cepstral mean subtraction with the mean obtained from the corresponding data set. In the following, we aim at estimating the corresponding clean envelope $c_\ell^{\hat{H}}(m)$ on basis of the preliminary denoised coefficients $c_\ell^H(m)$ from (11). Next, we provide our method to obtain a codebook for clean spectral envelopes, which is the backbone for the first three classification-based approaches as depicted in Fig. 3.

B. Codebook

The codebook $\mathcal{C} = \{\tilde{\boldsymbol{c}}_i^H\}$ consists of N_S envelope templates obtained from clean speech. Each template is represented by an N-dimensional vector of cepstral coefficients $\tilde{\boldsymbol{c}}_i^H = [\tilde{c}_i^H(1), \ldots, \tilde{c}_i^H(m), \ldots, \tilde{c}_i^H(N)]^{\mathrm{T}}$. Each entry of the codebook is representing a hidden state of the HMM, which is indexed by $i \in \{1, 2, \ldots, N_S\}$. We utilize the unsupervised Linde-Buzo-Gray (LBG) algorithm [36] to generate the codebook. We use an unsupervised method, since we are not interested in specific labels like, e.g., phonemes, but to obtain a good representation of many different envelopes. For training the codebook, any clean speech database is suitable. We use zero-mean clean speech envelope features (see Section IV-A) from frames identified by a simple energy threshold-based voice activity detection (VAD) as input to the LBG algorithm. The remainder of the clean speech training material is assigned an extra index $i = 0$, denoting non-speech frames, and is represented by an all-zero vector $\tilde{\boldsymbol{c}}_0^H = \boldsymbol{0}$ in the codebook. Accordingly, there are $N_S + 1$ states indexed by $i, j \in \mathcal{S} = \{0, 1, \ldots, N_S\}$. These states are to be estimated, e.g., by the HMM, which is introduced in the next subsection.

C. Hidden Markov Model

For the first two proposed approaches we will utilize a continuous density HMM to find a sequence of the hidden states $s_1, s_2, \ldots, s_\ell$, with $\lambda = \{\boldsymbol{\pi}, \boldsymbol{A}, b_j(\boldsymbol{x})\}$ being the set of parameters defining the HMM. Here, $\boldsymbol{\pi} = \{\pi_i\}$ denotes the initial state probability vector, $\boldsymbol{A}$ the state transition probability matrix with entries $a_{j,i} = \mathrm{P}(s_\ell = i|s_{\ell-1} = j)$ representing the probability to go from state $j \in \mathcal{S}$ into state $i \in \mathcal{S}$, and $b_i(\boldsymbol{x})$ the corresponding continuous emission probability density function for each hidden state. An observation is defined as $\boldsymbol{o}_\ell = \left[c_\ell^H(1), \ldots, c_\ell^H(N)\right]$, with $\boldsymbol{O}_\ell = \boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_\ell$ being a sequence of observations. The posterior distribution of the state probabilites given all the observations up to the current frame ℓ, $\mathrm{P}(s_\ell = i|\boldsymbol{O}_\ell)$, is obtained by applying the forward algorithm [37] as

$$\alpha_\ell(i) = b_i(\boldsymbol{x} = \boldsymbol{o}_\ell) \cdot \sum_{j \in \mathcal{S}} a_{j,i} \cdot \alpha_{\ell-1}(j), \tag{12}$$

followed by a normalization

$$\mathrm{P}(s_\ell = i|\boldsymbol{O}_\ell) = \frac{\alpha_\ell(i)}{\sum_{j \in \mathcal{S}} \alpha_\ell(j)}. \tag{13}$$

The first frame is initialized with $\alpha_1(i) = \pi_i \cdot b_i(\boldsymbol{x} = \boldsymbol{o}_1)$. In order to stay capable of realtime processing, we use the forward algorithm instead of, e.g., the forward-backward algorithm which would calculate the posterior distribution with even higher precision.

Having obtained the posterior distribution, we calculate the MMSE estimate

$$c_\ell^{\hat{H}}(m) = \sum_{i \in \mathcal{S}} \mathrm{P}(s_\ell = i|\boldsymbol{O}_\ell) \cdot \tilde{c}_i^H(m), \tag{14}$$

which represents a weighted average over all entries in the clean envelope codebook according to their respective probabilities. Alternatively, we use the maximum a posteriori (MAP) estimate

$$c_\ell^{\hat{H}}(m) = \tilde{c}_{i_\ell^*}^H(m) \tag{15}$$

with

$$i_\ell^* = \arg\max_{i \in \mathcal{S}} \alpha_\ell(i), \tag{16}$$

which simply selects the envelope with the highest posterior probability from the codebook. Here, the normalization from

https://doi.org/10.24355/dbbs.084-202102121510-0

(13) can be omitted, since it does not influence the $\arg\max$ operator. Note that for numerical stability we implemented our algorithms in the logarithmic domain.

The resulting zero-mean estimate of the clean envelope is required to maintain the channel properties, which is resolved by adding the corresponding cepstral mean. Finally, we calculate the spectral representation of the envelope as depicted by the SC blocks in Fig. 3. To accomplish this, we transform the estimated envelope back into N LPC coefficients by applying the following formula [35]:

$$\hat{a}_\ell(m) = c_\ell^{\hat{H}}(m) - \frac{1}{m}\sum_{\mu=0}^{m-1}\left[(m-\mu)\cdot c_\ell^{\hat{H}}(m-\mu)\cdot\hat{a}_\ell(\mu)\right] \tag{17}$$

for $1 \leq m \leq N$. Its spectral representation $|\hat{H}_\ell(k)|$ is received by first applying a K-point DFT to the LPC coefficients, padded with $K-N-1$ zeros. This results in

$$\left(\hat{A}_\ell(k)\right)_{k=0}^{K-1} = \text{DFT}\left\{(0, \hat{a}_\ell(1), \ldots, \hat{a}_\ell(N), 0, \ldots, 0)\right\}, \tag{18}$$

followed by

$$|\hat{H}_\ell(k)| = \frac{1}{|1-\hat{A}_\ell(k)|}. \tag{19}$$

The initial state distribution vector $\boldsymbol{\pi}$ is assuming a uniform distribution (and is therefore not effective in Fig. 3), while the required state transition matrix $\boldsymbol{A}$ is generated by counting transitions between the states in the clean training material followed by a normalization to calculate the conditional probabilities.

In the following, we will present two different methods to model the observations in order to obtain emission probabilities $b_i(\boldsymbol{x})$ by using either GMMs, or a classification DNN with prior division. We further investigate using the posterior distribution from a classification DNN directly, or a regression DNN (Section IV-E), directly estimating clean coefficients from the preliminary denoised observations. We then provide a generic description of the DNN training mechanisms in Section IV-F and will finally show, how these four CEE schemes can be combined with CEM, if desired.

D. HMM With GMM or Classification DNNs

Now, with the hidden states obtained from the LBG algorithm as described in Section IV-B, we generate quite some training material that represents typical observations for the HMM. To accomplish that, we simulate various SNR and noise conditions with the same clean speech data that has been used to retrieve the hidden states. This noisy speech data is subsequently processed by a preliminary noise reduction scheme running with the same parameterization as will be used for testing. It is followed by source-filter decomposition via LPC analysis, where only the envelope is used further for the GMM/DNN training. During this process it is important to keep track of the corresponding hidden state for each processed frame by knowing the quantization index i of its equivalent in the clean envelope codebook $\mathcal{C}$. This is required in order to obtain an assignment between a clean envelope and all its corresponding denoised observations. With the aid of this information we are able to train models which represent the denoised observations for each of the states. We introduce GMMs and DNNs as such models in the following two subsections and also show how to replace the HMM completely by a DNN in the third.

*1) GMM-Based HMM (**GMM-HMM**):* For each state i and its corresponding training material (representing observations) we use the expectation maximization (EM) algorithm [38] to train all parameters of a GMM with G modes, separately. The GMM is representing statistics of the preliminary denoised envelope observations which is later on mapped to a hidden state. In that fashion we receive the required models for the emission probabilities $b_i(\boldsymbol{x})$, $i \in \mathcal{S}$.

The observation probabilities for a certain input $b_i(\boldsymbol{x} = \boldsymbol{o}_\ell)$ are obtained by evaluating each GMM as follows

$$b_i(\boldsymbol{x} = \boldsymbol{o}_\ell) = \sum_{g\in\mathcal{G}} c_{i,g}\cdot\mathcal{N}(\boldsymbol{x} = \boldsymbol{o}_\ell; \boldsymbol{\mu}_{i,g}, \boldsymbol{\Sigma}_{i,g}), \tag{20}$$

with $g \in \mathcal{G} = \{1, \ldots, G\}$ being the mode index, weights $c_{i,g}$ constrained to $\sum_{g\in\mathcal{G}} c_{i,g} = 1$, $\boldsymbol{\mu}_{i,g}$ as mean vectors, and $\boldsymbol{\Sigma}_{i,g}$ being the (in our case diagonal) covariance matrix for each corresponding mode g and state i. It plugs directly into (12) and is representing the GMM block in the upper path (S_1 in upper position) in Fig. 3.

*2) DNN-Based HMM (**C-DNN-HMM**):* An alternative to GMMs as observation models is a feedforward DNN trained as classifier. The output of the classification DNN, the posterior probabilities for each of the hidden states given the current observation, is defined as $\mathrm{P}(s_\ell = i|\boldsymbol{x} = \boldsymbol{o}_\ell)$. To use the output of the DNN in the HMM framework (12) (Fig. 3, second path from top, S_1 in center position, C-DNN block, and S_2 in upper position) we actually need to divide it by the prior state probability to obtain the likelihood as

$$b_i(\boldsymbol{x} = \boldsymbol{o}_\ell) = \mathrm{p}(\boldsymbol{x} = \boldsymbol{o}_\ell|s_\ell = i) \propto \frac{\mathrm{P}(s_\ell = i|\boldsymbol{x} = \boldsymbol{o}_\ell)}{\mathrm{P}(s = i)}, \tag{21}$$

with $\mathrm{P}(s = i) = \pi_i$. We omit the evidence $\mathrm{p}(\boldsymbol{x})$ as it has only a normalizing function.

*3) DNN Without HMM (**C-DNN**):* A further option to obtain posterior state probabilities is to use the output of a classification DNN directly (Fig. 3, third path from top, S_1 in center position, C-DNN block, and S_2 in lower position) and to omit the HMM framework, thereby losing the advantage of the temporal modeling from the HMM. Since we can understand the output of the DNN as $\mathrm{P}(s_\ell = i|\boldsymbol{x} = \boldsymbol{o}_\ell)$, we can use it directly for either MMSE estimation as shown in (14) or MAP estimation in (16), where it is necessary to replace $\alpha_\ell(i)$ with the DNN output.

Next, we introduce a solution that is independent of a codebook or an HMM and estimates the clean envelope representing coefficients directly.

*E. Regression DNN (**R-DNN**)*

Instead of using DNNs as a classifier, it is also possible to directly estimate enhanced coefficients $c_\ell^{\hat{H}}(m)$ from a denoised input vector by means of regression. The output plugs directly

https://doi.org/10.24355/dbbs.084-202102121510-0

into the spectral conversion (SC) block in Fig. 3 and renders a codebook needless (Fig. 3, lowest path, S_1 in lower position, R-DNN block). In this particular case, the temporal context is also lost, unless the input layer of the DNN supports multiple input frames.

The coming subsection gives a brief overview of the training procedures required for the introduced DNN approaches.

F. DNN Training

To maintain comparability, we use the same zero-mean input features for the DNN as for the GMMs, and for regression also zero-mean targets. The number of nodes for the input layer is corresponding to the feature vector dimension N and the number of nodes for the output layer corresponding either to the amount of classes $N_S + 1$, or also to the feature dimension N (regression training). We understand hidden layers as every layer between the input and the output layer and their number is N_H, where each hidden layer has N_N nodes. The initialization of the network's parameter set, comprising the weights and biases, is done as proposed by Glorot *et al.* in [39]. In order to obtain posterior class probabilities we use the negative log-likelihood (NLL) error criterion during training with the backpropagation algorithm [40] and a softmax output layer. The difference for a regression-based DNN is mainly the final layer, which is a linear output layer in this case. Also, the used error criterion during the training is the mean squared error (MSE) instead of NLL. As activation functions in the other layers we employ sigmoid functions or rectified linear units (ReLUs). The latter are resolving the vanishing gradient issue [41], known to occur with sigmoid functions. After network initialization, the training material is randomly assigned to batches containing L input frames each. Then, according to the error criterion, the gradients of the loss function between the outputs of the network and the corresponding targets are calculated for each batch, and are subsequently backpropagated through the network. The deltas of the parameters are accumulated and finally the network's weights and biases are updated. We train each network with $L = 1024$ samples (frames) per batch and a fixed learning rate of $\eta = 0.001$ for 100 epochs. Finally, we select the model with the best performance on the development set for speech active frames (H_1), as experiments with adaptive learning rate decay have shown to perform only as good as but not better. Also, the investigation of L2 regularization did not lead to improvement, even worse, we could witness some configurations, where the networks deteriorate and classify every input as speech inactive (H_0).

Next, we provide instructions on how to apply or combine some of the introduced approaches.

G. Applications With CEM

The CEE scheme can be combined with CEM in two different ways: A parallel structure, where CEM and CEE are applied simultaneously, meaning that the CEE block from Fig. 2 is placed into the lower path of the LPC analysis in Fig. 1, or a serial structure where the systems from Fig. 1 and Fig. 2 are cascaded in either way. Here, cascading means that the output of the first system (being either $\hat{S}_\ell^{\mathrm{CEM}}(k)$ or $\hat{S}_\ell^{\mathrm{CEE}}(k)$) is used as input for the LPC analysis block of the second system, thereby replacing $\bar{Y}_\ell(k)$. Hence, the preliminary noise reduction of the second system is omitted and the noise power estimate $\sigma_\ell^D(k)^2$ of the first system is used throughout. The final gain function is also applied to the original microphone signal $Y_\ell(k)$.

V. Experimental Setup

A. CEM and DD Baselines (*CEM$_{SI}$* and *DD*)

As we have already shown in [8], our baseline CEM algorithm outperforms several state of the art *a priori* SNR estimation algorithms. As motivated before, our experiments aim at further enhancing the CEM algorithm by employing our various envelope estimators, and compare the new approach to the speaker-independent CEM baseline (**CEM$_{SI}$**) and also the DD estimator (**DD**) which is parameterized with $\xi_{\min} = -15$ dB and $\beta_{\mathrm{DD}} = 0.975$. For a detailed setup of the training for the **CEM$_{SI}$** approach, we kindly refer to [8].

B. Ideal Ratio Mask Baseline (*IRM*)

In addition, we simulate a data-driven baseline using a feed-forward neural network that predicts the ideal ratio mask (IRM). This baseline DNN has 2,364,545 parameters and is mostly in line with Wang's work ([42] and [20]). We use non-redundant amplitude features compressed by the natural logarithm as input features, while the **IRM** targets are calculated as

$$G_\ell^{\mathrm{IRM}}(k) = \left(\frac{|S_\ell(k)|^2}{|S_\ell(k)|^2 + |D_\ell(k)|^2} \right)^\beta, \tag{22}$$

with $\beta = 1.0$. By interpreting the **IRM** as a gain function, we are able to integrate this baseline into our evaluation methodology (we require separately processed speech and noise *components*, as will be outlined at the end of this section). As some of our introduced approaches are based on the **CEM$_{SI}$** baseline, and thus indirectly on the **DD** baseline, we will first report the performance of our approaches w.r.t. the two baselines for our development process. However, for the final evaluation on the test data, we will also compare our approaches with the data-driven **IRM** baseline.

C. Databases and Preprocessing

We evaluate the algorithms in a noise reduction framework and analyze the performance in a total of 318 different conditions, embracing six different SNRs from -5 dB up to 20 dB in steps of 5 dB, and 53 different noise files where we use all 20 files from the QUT [43] database and 33 out of 38 files from the ETSI [44] database. Among them we find noise types such as babble, car, street, aircraft, train, work, and more. We leave out the male single voice distractor noise file and hold out four further noise files from the ETSI database for an extra test set with noise files which have not been seen during training. We split each noise file into three non-overlapping parts, where 60% are used for training, 20% for the development set, and another 20% for the test set. As clean speech databases we utilize the TIMIT [45] and also the NTT super wideband database [46] (American and British English only), both downsampled to 8 kHz.

https://doi.org/10.24355/dbbs.084-202102121510-0

The designated training set of the TIMIT database is used for training while the test set is used as development set and the NTT database is used for testing only. We decided to utilize the databases in that way since the training process requires a lot of data which the TIMIT database delivers and also we are able to show performance across different databases. For evaluation of the training and development set with our **CEM$_{\mathbf{SI}}$** approach, we use one speaker-independent codebook based on the NTT database and for the test sets we use the speaker-independent codebooks as obtained in [8].

The various SNR conditions are obtained by measuring and adjusting the levels of the randomly selected noise portions and clean speech files after ITU-T P.56 [47], followed by their superposition. The framing (analysis and also overlap-add synthesis) is done with a periodic square root Hann window and a 50% frame shift, where one frame embraces $K = 256$ samples. The LPC analysis calculates $N{=}10$ LPC coefficients. Furthermore, we conduct the DNN training with the `Torch` toolkit [48] on `CUDA`-capable GPUs.

D. Instrumental Quality Assessment

For the quality assessment we employ the white-box approach [49], which means that we apply the calculated gains $G_\ell(k)$ not only to the microphone signal $Y_\ell(k)$ to obtain the clean speech estimate $\hat{S}_\ell(k)$, but also to the components $S_\ell(k)$ and $D_\ell(k)$, separately. We refer to the resulting entities after IDFT and overlap-add as the *filtered* clean speech component $\tilde{s}(n)$ and the *filtered* noise component $\tilde{d}(n)$, respectively. As instrumental measures we use the segmental noise attenuation (NA$_{\text{seg}}$) [50] which is calculated as

$$\text{NA}_{\text{seg}} = 10\log_{10}\left[\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\text{NA}(\ell)\right], \tag{23}$$

where

$$\text{NA}(\ell) = \frac{\sum_{\nu=0}^{N-1} d(\nu+\ell N)^2}{\sum_{\nu=0}^{N-1}\tilde{d}(\nu+\ell N+\Delta)^2},$$

with $\ell \in \mathcal{L}$ defining a segment of $N = 256$ samples, Δ being the compensation term for potential delay due to filtering, and a normalizing factor $\frac{1}{|\mathcal{L}|}$, taking into account the number of all frames. Furthermore, we also evaluate the delta SNR as

$$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}. \tag{24}$$

Here, SNR$_{\text{in}}$ depicts the SNR of the clean speech and noise component while SNR$_{\text{out}}$ depicts the SNR of the *filtered* speech and noise components, after processing. This measure allows to draw conclusions on the actual improvement of the SNR, since a high noise attenuation might also affect the speech component.

We also employ the PESQ score (mean opinion score, listening quality objective (MOS-LQO)) [51], [52], on the *filtered* clean speech component $\tilde{s}(n)$ with $s(n)$ as reference. Thereby, we are able to evaluate the noise and also the speech *components* separately. We do *not* measure PESQ on the enhanced signal $\hat{s}(n)$, since PESQ has not been validated for artifacts caused by noise reduction techniques. In line with P.1100 [53, Sect. 8] and using [49] to obtain the processed clean speech component, we instead measure the distortion of the clean speech component, thereby being also compliant to the intended use case of P.862 [51]. Since PESQ is somewhat level-agnostic and thus not penalizing broadband attenuation of a signal, it is important to report the ΔSNR. This allows to draw conclusions on speech attenuation which would not be possible otherwise.

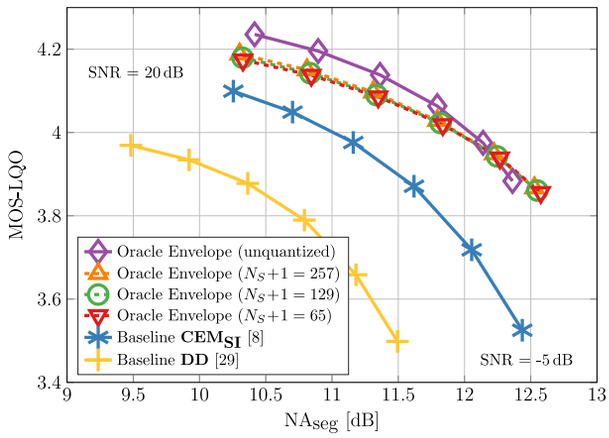


Fig. 4. Evaluation of the *speech component* MOS-LQO and segmental NA for all noise and SNR conditions of the unquantized and quantized **oracle experiments** and the **two baseline estimators** showing the potential of the proposed method on the **development set**.

In order to assess the intelligibility of the enhanced speech, we employ the short-time objective intelligibility measure (STOI) [54]. STOI is an intrusive metric that is operating on the clean speech signal $s(n)$ which serves as a reference and the enhanced signal $\hat{s}(n)$. This metric provides values in the range $[0, 1]$, where high values represent high intelligibility.

VI. Simulations and Discussion

A. Solo: Cepstral Envelope Estimation (CEE)

Number of HMM States: At first, we perform two different oracle experiments in order to analyze the potential of our approach and to figure out how many states are providing good performance. In Fig. 4, we evaluate MOS-LQO by PESQ and also NA$_{\text{seg}}$, both measured on the separately filtered components. Here, each marker depicts a certain SNR condition, with -5 dB in the lower right and 20 dB in the upper left corner. The solid purple plot (with diamond markers) shows the performance of the proposed method when instead of the applied CEE (see Fig. 2, grey box), the oracle envelope from the clean speech is injected and mixed with the denoised residual signal (referred to as Oracle Envelope). Accordingly, this plot depicts the upper performance limit of the CEE technique in our noise reduction framework. Now, the first choice we need to make is on the amount of states the HMM should be able to estimate. Therefore, we train three different codebooks (see Section IV-B) for $N_S \in \{64, 128, 256\}$ with the LBG algorithm [36] on the extracted envelopes of the TIMIT training set. Subsequently, we run our framework, again replacing the CEE block by quantizing the oracle envelopes obtained from the corresponding clean speech files with our trained LBG codebooks (three dashed lines, triangle and circle markers). Comparing both oracle experiments to the **DD** (solid yellow line, plus markers) and **CEM$_{\mathbf{SI}}$** (solid blue line, asterisk markers) baselines, shows that there

https://doi.org/10.24355/dbbs.084-202102121510-0

TABLE I
ANALYSIS OF THE **GMM-HMM** APPROACH WITH $N_S + 1 = 65$ STATES, G BEING THE NUMBER OF MODES: **POSTERIOR STATE PROBABILITY ACCURACY** DELIVERED BY THE HMM. SPEECH ACTIVE (H_1) AND INACTIVE (H_0) FRAMES ARE EVALUATED SEPARATELY

	$H_0 \cup H_1$			H_0			H_1		
G	4	8	16	4	8	16	4	8	16
Train	0.4358	0.4421	0.4460	0.6376	0.6455	0.6433	0.3648	0.3705	0.3766
Dev	0.4286	0.4352	**0.4376**	0.6332	**0.6454**	0.6415	0.3587	0.3635	**0.3680**
Test	0.4993	0.5034	0.5014	0.6415	0.6474	0.6397	0.3419	0.3440	0.3484

is good potential of the approach, especially in terms of speech component quality. One can also see that the quantization causes a slightly higher NA_{seg} in the lower SNR conditions compared to the oracle envelope, where in the other SNR conditions it is more a loss in speech component quality only. The three dashed lines representing the different quantization levels show a very similar performance with only a slight preference for the larger codebooks. However, since it is only a marginal benefit, we decide to use $N_S + 1 = 64 + 1$, as the trade-off between lower complexity and higher quality clearly favors the former in this case.

Number of GMM Modes: Next, we investigate the number of modes G which represent the denoised observations. Therefore, we train GMMs with $G \in \{4, 8, 16\}$ and evaluate the posterior state probabilities of the HMM by measuring the accuracy. The results are shown in Table I and are depicted for speech active (H_1), speech inactive (H_0), and both kinds of frames together ($H_0 \cup H_1$). The H_0/H_1 distinction is performed by a simple VAD on the clean speech material with a dynamic threshold which tests if a frame's energy is above the average frame energy of the corresponding file. The rationale behind this is that the prior distribution of the state representing speech inactive frames differs between the three sets, being roughly 25% for the training and development set, and 50% for the test set. This, if only regarding the accuracy of all frames jointly, would raise questions as to why the accuracy on the test set is higher than on the training and development set. Considering both classes separately gives a more consistent view on the performance, showing an expectedly higher accuracy with increasing number of modes on the speech active frames. The gain is rather small compared to the rising complexity with increasing G, making us comfortable with the choice of $G = 16$ (grey-shaded), delivering the best accuracy for speech active frames on the development set, without exploring the effects of more modes which we assume would lead to overfitting at some point and also to a lack of training data. Fortunately, this coincides with the best H_1 performance on the test set as well, which is not taken for granted.

GMM-HMM Envelope Estimation: Thus, having found a suitable configuration we evaluate the performance of the **GMM-HMM** approach with $N_S + 1 = 65$ states each represented by $G = 16$ modes with either MAP (16) or MMSE (14) estimation of the clean envelope in Fig. 5. On top, the unquantized and also quantized oracle experiments with $G = 16$ are shown. Compared to the **DD** baseline (solid yellow line, plus markers) the MAP approach (dashed green line, square markers) is able to show consistent improvement in terms of both measures, MOS-LQO and NA_{seg}. Especially the low-SNR conditions benefit from the enhanced envelope in terms of speech component

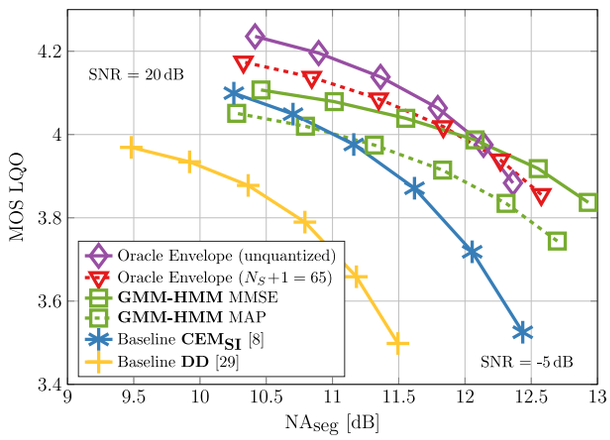


Fig. 5. Evaluation of the *speech component* MOS-LQO and segmental NA for all noise and SNR conditions of the **two optimized GMM-HMM approaches** using MAP and MMSE ($N_S + 1 = 65, G = 16$) compared to the two corresponding **oracle experiments** and **baseline estimators** showing the performance of the optimized **GMM-HMM** approaches on the **development set**.

TABLE II
ABERRATION OF PARAMETERS FOR SEVERAL NUMBERS OF HIDDEN LAYERS N_H WITH NUMBER OF NODES N_N IN EACH LAYER COMPARED TO THE DESIGNATED GMM CONFIGURATION WITH 21840 PARAMETERS

N_H	1	2	3	4	5	6
N_N	286	114	86	73	64	58
# Parameters	21801	21839	21565	21819	21569	21583
Aberration %	0.18	0.01	1.26	0.10	1.24	1.18

quality. The proposed approach also exceeds the $\textbf{CEM}_{\textbf{SI}}$ baseline (solid blue line, asterisk markers) in the SNR conditions from -5 dB up to 10 dB quite clearly. Only the two best SNR conditions enable the CEM approach to obtain better speech component quality, which gives hope that a combination of both approaches might be able to mitigate the drawbacks of either method. When evaluated against the corresponding oracle envelope experiment (dashed red line, triangle markers) a more or less constant gap of around 0.05 MOS points remains. To circumvent the limitation of using a single entry of the codebook only, as done by the MAP estimation, we also calculate the MMSE estimate (solid green line, square markers), allowing us to consistently exceed the performance of the MAP approach by up to 0.09 MOS points for the -5 dB SNR condition. Even the oracle envelope experiment can be outperformed in terms of NA_{seg}, however, with a slightly lower MOS-LQO. This depicts nicely the benefit of the MMSE over the MAP estimate, being able to exploit the codebook space to a larger extent. The experiment using the unquantized oracle envelope performs clearly better than the **GMM-HMM** with MMSE estimation in the 20 dB to 10 dB SNR conditions, while in the remaining SNR conditions the MMSE approach obtains a much higher NA_{seg} which might be caused by a less accurate state estimation due to the SNR, being reflected by the lower MOS-LQO values.

C-DNN Envelope Estimation Approaches: The GMMs with $G = 16$ and $N_S + 1 = 65$ embrace a total of 21,840 parameters, which we target also for the training of DNNs to ensure a fair comparison. In Table II we depict several basic network configurations with up to six hidden layers, trying to keep a comparable amount of parameters as used for GMM training and we

https://doi.org/10.24355/dbbs.084-202102121510-0

TABLE III
EVALUATION OF **VARIOUS C-DNN TRAININGS** WITH COMPARABLE AMOUNT OF PARAMETERS RELATED TO THE BEST **GMM** CONFIGURATION IN TERMS OF THE **POSTERIOR STATE PROBABILITY ACCURACY** DELIVERED BY THE RESPECTIVE DNN. SPEECH ACTIVE (H_1) AND INACTIVE (H_0) FRAMES ARE EVALUATED SEPARATELY. THE EPOCH #E OF THE BEST PERFORMING NETWORK WITH RESPECT TO ACCURACY ON H_1 ON THE DEVELOPMENT SET IS ALSO REPORTED

				Training Set			Development Set			Test Set		
N_H	N_N	Activation	#E	$H_0 \cup H_1$	H_0	H_1	$H_0 \cup H_1$	H_0	H_1	$H_0 \cup H_1$	H_0	H_1
1	286	Sigmoid	55	0.5573	0.8788	0.4442	0.5452	0.8762	0.4321	0.6309	0.8529	0.3851
		ReLU	99	0.5552	0.8786	0.4415	0.5435	0.8768	0.4296	0.6312	0.8525	0.3863
2	114	Sigmoid	100	0.5596	0.8816	0.4463	0.5461	0.8775	0.4329	0.6337	0.8565	0.3871
		ReLU	6	0.5571	0.8813	0.4430	0.5439	0.8779	0.4299	0.6312	0.8514	0.3874
3	86	Sigmoid	68	0.5605	0.8803	0.4481	0.5467	0.8758	0.4344	0.6364	0.8592	0.3898
		ReLU	100	0.5588	0.8840	0.4444	0.5450	0.8801	0.4306	0.6351	0.8621	0.3838
4	73	Sigmoid	50	0.5612	0.8813	0.4485	**0.5472**	0.8757	0.4349	0.6385	0.8642	0.3887
		ReLU	55	0.5583	0.8863	0.4429	0.5451	**0.8823**	0.4299	0.6372	0.8668	0.3830
5	64	Sigmoid	89	0.5616	0.8809	0.4493	**0.5472**	0.8759	0.4349	0.6360	0.8616	0.3864
		ReLU	93	0.5581	0.8819	0.4443	0.5444	0.8770	0.4309	0.6358	0.8602	0.3873
6	58	Sigmoid	33	0.5608	0.8782	0.4492	0.5469	0.8734	**0.4354**	0.6367	0.8617	0.3876
		ReLU	41	0.5577	0.8828	0.4433	0.5445	0.8786	0.4304	0.6353	0.8607	0.3857

also depict the aberration of parameters in percent. Thereby, we make sure that we do not use more parameters than the GMM baseline does. With this setup we are able to analyze the influence of depth independently. Following, in Table III we present the posterior state accuracies of the various **C-DNN** configurations which we have examined. In general, there is to say that the differences between the configurations are rather small so that no network performs significantly better than any other. Judging from the development set, the networks with sigmoid activation obtain slightly better accuracies on speech active frames than the ones with ReLUs. Another observation is that with growing depth we can see a slight but steady increase of the accuracy on the H_1 frames of the development set when the sigmoid function is employed. For the subsequent **C-DNN-HMM** and also **C-DNN** approaches, we use the network with $N_H = 6$, $N_N = 58$, and sigmoid activation, as it performs best on speech active frames on the development set (grey-shaded). Note that it just does not match the best results on the test set which the network with $N_H = 3$ and $N_N = 86$ yields. When compared to the best **GMM-HMM** result in Table I (43.8%), the superiority of the DNN (54.7%) becomes obvious, as the accuracy gain on the development set ($H_0 \cup H_1$) is better than 10% absolute, and also on the speech active frames the accuracy increases by more than 6%. For the test set, the overall accuracy is more than 13% higher, while the gain for speech active frames of the test set melts down to about 4%.

R-DNN Envelope Estimation: Results of the second training process for the **R-DNN** are shown in Table IV. Again, we made sure that the amount of parameters relates closely to the best performing **GMM-HMM** configuration. The DNNs trained with sigmoid activation function slightly outperform the ones with ReLU activation function, as before. However, the latter tend to converge a bit faster for some topologies but with a higher loss. In general, the differences across all configurations are rather marginal. Nevertheless, we find the best configuration for $N_H = 6$ and $N_N = 58$ combined with the sigmoid activation function (grey-shaded). This is the same configuration as we found to be optimal for the **C-DNN** approaches. Also, this network shows only second best performance on the test set.

TABLE IV
EVALUATION OF **VARIOUS R-DNN TRAININGS** WITH COMPARABLE AMOUNT OF PARAMETERS RELATED TO THE BEST GMM CONFIGURATION IN TERMS OF THE **MSE LOSS**. THE EPOCH #E OF THE BEST PERFORMING NETWORK WITH RESPECT TO THE MINIMAL MSE LOSS ON THE DEVELOPMENT SET IS ALSO REPORTED

				Training Set	Development Set	Test Set
N_H	N_N	Activation	#E	MSE loss		
1	286	Sigmoid	100	0.0480	0.0495	0.0506
		ReLU	76	0.0482	0.0498	0.0508
2	114	Sigmoid	100	0.0468	0.0485	0.0504
		ReLU	100	0.0471	0.0487	0.0507
3	86	Sigmoid	100	0.0465	0.0483	0.0515
		ReLU	95	0.0468	0.0485	0.0502
4	73	Sigmoid	100	0.0464	0.0481	0.0501
		ReLU	100	0.0468	0.0485	0.0501
5	64	Sigmoid	93	0.0464	0.0481	0.0496
		ReLU	50	0.0468	0.0485	0.0498
6	58	Sigmoid	91	0.0463	**0.0480**	0.0498
		ReLU	59	0.0468	0.0485	0.0499

All Approaches: Now, we evaluate the performance of the optimal networks in our system for the MAP estimation, as shown in Fig. 6. Comparing the **GMM-HMM** approach (dashed green line, square markers) to the **C-DNN-HMM** configuration (dashed orange line, triangle markers), results in an unchanged performance, which is surprising, since the accuracy of the **C-DNN** alone is significantly higher. A gain is seen, however, for the **C-DNN** (dashed orange line, circle markers), where the HMM is omitted and the posterior distribution of the network is used directly. An analysis of the state posterior distribution accuracy on the development set shows that the reported 54.7% ($H_0 \cup H_1$) and 43.5% (H_1) of the **C-DNN** (both Table III) correspond to only 45.2% ($H_0 \cup H_1$) and 38.0% (H_1) for the **C-DNN-HMM** approach, which is still higher by more than 1% compared to the **GMM-HMM** method (cf. Table I). However, this latter only small accuracy improvement explains the comparable performance of **C-DNN-HMM** and **GMM-HMM** in Fig. 6. The **C-DNN** consistently outperforms the two HMM-based systems in both quality dimensions by up to 0.05 MOS

https://doi.org/10.24355/dbbs.084-202102121510-0

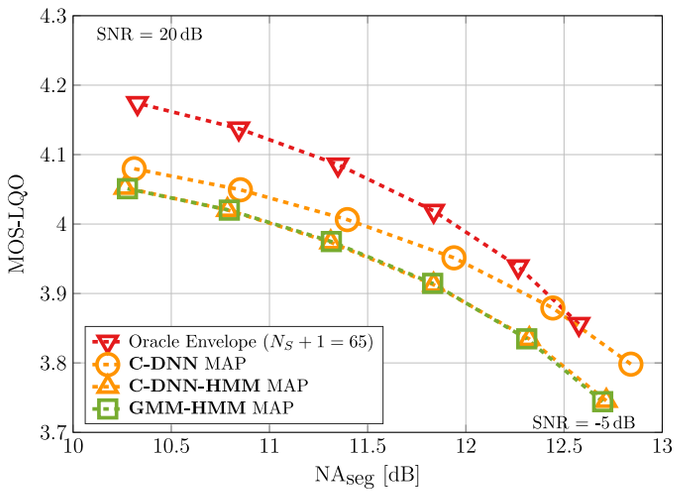


Fig. 6. Comparing the performance of the **GMM-HMM** system ($N_S + 1 = 65$, $G = 16$) and the various **DNN-supported approaches** with **MAP** estimation in terms of the speech component quality measured by MOS-LQO and NA$_{seg}$ on the **development set**. The upper limit is depicted by the respective oracle experiment.

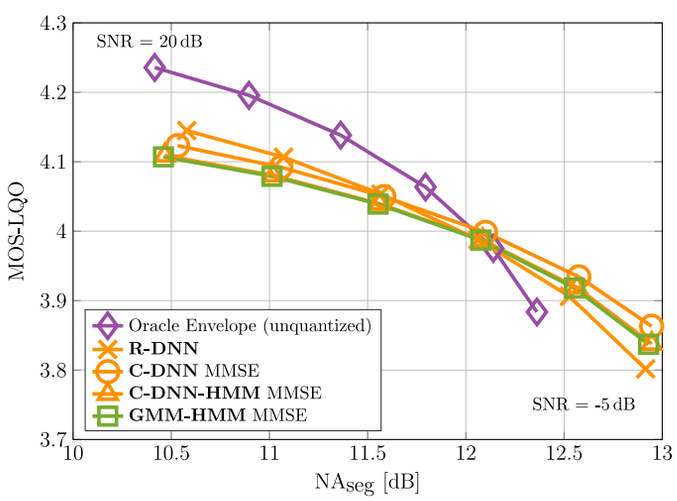


Fig. 7. Comparing the performance of the **GMM-HMM** system ($N_S + 1 = 65$, $G = 16$) and the **various DNN-supported approaches** with **MMSE** estimation in terms of the speech component quality measured by MOS-LQO and NA$_{seg}$ on the **development set**. The upper limit is depicted by the respective oracle experiment.

points and 0.1 dB NA (−5 dB SNR condition), showing improved performance especially in the low-SNR conditions. This indicates that the HMM seems to be a limiting factor here, which could be caused by the temporal context, since it is the remaining factor that is able to overrule the network's decision.

The results for the MMSE estimation are reported in Fig. 7. Again, we can see that replacing the GMMs by a DNN (solid green line, square markers: **GMM-HMM** vs. solid orange line, triangle markers: **C-DNN-HMM**) has very little effect due to the limiting HMM. The performance of the **C-DNN** (solid orange line, circle markers) again shows consistent improvement over the HMM results, which indicates that the overall estimation of the posterior probability distribution is more accurate. Given the 10% accuracy improvement of **C-DNN** vs. **GMM-HMM**, and the 56.2% accuracy improvement of the oracle vs.

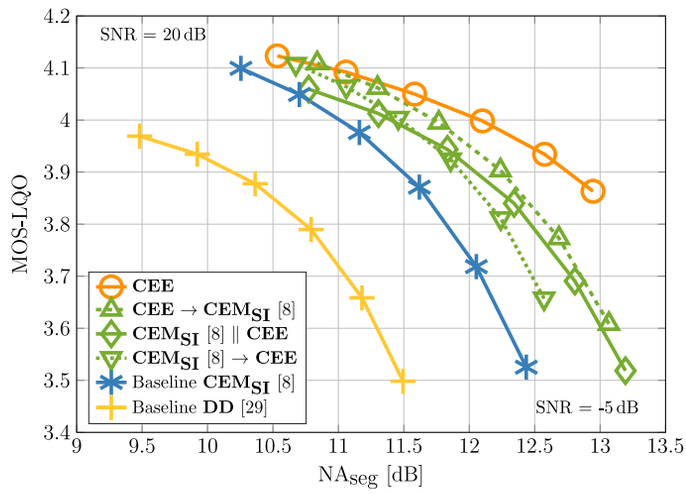


Fig. 8. Comparing the performance of the **CEE** system, the **baselines CEM$_{SI}$** and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech component quality measured by MOS-LQO and NA$_{seg}$ on the **development set**.

GMM-HMM, **C-DNN** performs better than expected. This is visible, e.g., in SNR = −5 dB, where its MOS-LQO is about half way between **GMM-HMM** and oracle, while it exceeds the oracle NA$_{seg}$ by more than 0.5 dB. Finally, the **R-DNN** (solid orange line, cross markers) shows an imbalanced behavior as it exceeds the performance of the **C-DNN** for the 15 and 20 dB SNR conditions but deteriorates with decreasing SNR. This results in the worst performance among the depicted methods for the two lowest SNR conditions. This is an interesting result as this shortcoming could not be observed for the classification DNNs. It could be due to the rather small amount of parameters, preventing the network to cover all SNR conditions equally as the regression task is more complex than classification. Consequently, we favor the **C-DNN** approach with MMSE estimation, as it performs best in the important low-SNR conditions. The approach still leaves space for improvement, especially for the higher SNR conditions, when compared to the oracle experiment.

B. Duo: CEM With Cepstral Envelope Estimation (CEE)

Having successfully identified the best performing envelope estimator, namely the **C-DNN** approach with MMSE estimation, which we will simply dub **CEE** in the following, we will now combine **CEE** with **CEM$_{SI}$** by replacing the preliminary denoised envelope in Fig. 1 (lower LPC analysis path, white box) with the proposed **C-DNN** cepstral envelope estimation method. This is referred to as parallel approach (symbol ‖). Alternatively, we will also investigate using either **CEM$_{SI}$** or the **C-DNN** approach as preliminary noise reduction for the other, referred to as serial approaches (symbol →).

1) Evaluation on the Development Set: The results are depicted in Fig. 8, where the **CEM$_{SI}$** (solid blue line, asterisk markers) benefits especially in the important low-SNR conditions from incorporating the **CEE** (solid orange line, circle markers) in a parallel manner (solid green line, diamond markers) by obtaining a higher NA$_{seg}$ while maintaining a

https://doi.org/10.24355/dbbs.084-202102121510-0

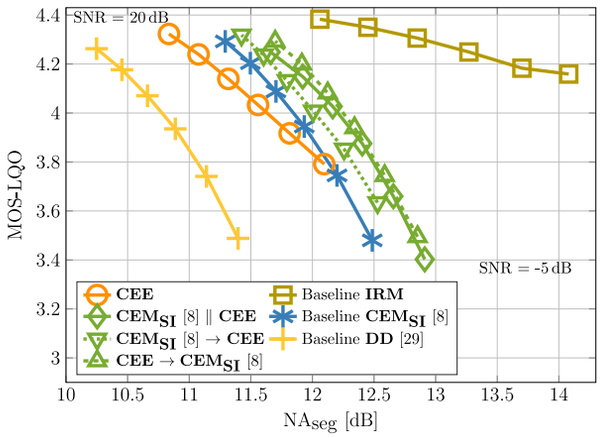


Fig. 9. Comparing the performance of the **CEE** system, the **baselines CEM$_{SI}$**, **IRM**, and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech component quality measured by MOS-LQO and NA$_{seg}$ on the **test set**.

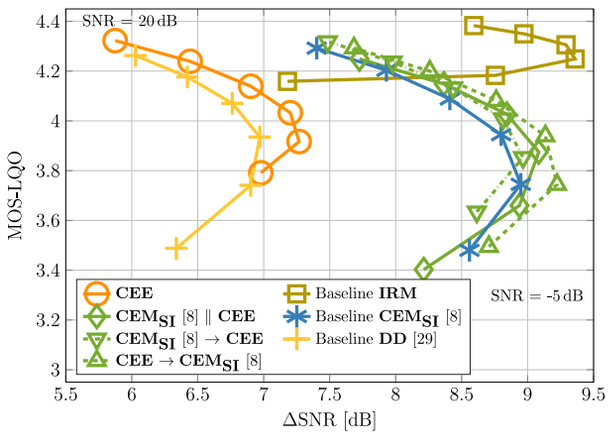


Fig. 10. Comparing the performance of the **CEE** system, the **baselines CEM$_{SI}$**, **IRM**, and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech component quality measured by MOS-LQO and **ΔSNR** on the **test set**.

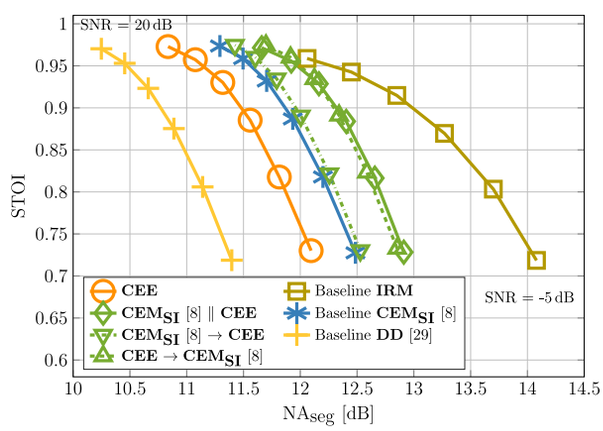


Fig. 11. Comparing the performance of the **CEE** system, the **baselines CEM$_{SI}$**, **IRM**, and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech intelligibility measured by **STOI** and NA$_{seg}$ on the **test set**.

comparable speech component quality. Also the serial approaches (green lines, triangle markers) both outperform the **CEM$_{SI}$** baseline consistently in both quality dimensions, gaining up to 0.63 dB higher NA$_{seg}$ and 0.13 MOS points. Applying **CEM$_{SI}$** first (dotted green line, inverted triangle markers) followed by the **CEE** yields a slightly higher speech component quality at the cost of a little less NA$_{seg}$ compared to the other serial setup (dashed green line, triangle markers). The **CEM$_{SI}$** approaches, solo and duo, have one important advantage over the solo **CEE** approach: They are able to restore harmonics and to suppress noise between them, where the latter is a shortcoming of all approaches which only estimate the envelope. However, we expected a more consistent improvement by applying both techniques in parallel and suspect that some mismatch between the enhanced excitation and envelope could prevent further improvement, which could be subject to future research. This mismatch seems to be eased by the sequential application of both approaches, where we manipulate one component of the estimated clean speech amplitude spectrum at a time.

2) Evaluation on the Test Set: Until now, all results and optimizations have been analyzed and taken out on the development set. In Fig. 9 we report the test set performance of the three baseline approaches, **DD** (solid yellow line, plus markers), **CEM$_{SI}$** (solid blue line, asterisk markers), and **IRM** (solid sand line, square markers). We also report on our best cepstral envelope estimator **C-DNN** with MMSE estimation, i.e., **CEE** solo (solid orange line, circle markers), and also in conjunction (green lines) with the **CEM$_{SI}$** baseline. When the solo **CEE** approach is applied, a consistent improvement of the speech component quality over the **DD** and **CEM$_{SI}$** baselines is obtained, but the NA$_{seg}$ now falls behind the **CEM$_{SI}$** method. This probably reflects the detriment of the **CEE** approach being a data-driven technique, since this was not the case on the development set. Interestingly, the two baseline approaches (**DD**, **CEM$_{SI}$**) yield lower PESQ scores on the development set than on the test set (compare Figs. 8 and 9). This is most likely due to the choice of two different databases which have quite different recording characteristics and settings. Thus, the one seems to be easier to be processed by noise reduction algorithms than the other. As the other approaches (green lines) are heavily influenced by the data-driven **CEE** approach, which has been trained on data stemming from the same database (but disjoint data sets) as the development set, the decreasing performance is quite expected when changing to a different database. However, the combination with **CEM$_{SI}$** seems to mitigate the drawback of the **CEE** approach caused by its data dependency to quite some extent. In parallel with the **CEM$_{SI}$** (solid green line, diamond markers) a gain of up to 0.4 dB NA$_{seg}$ can be obtained, resulting in a slight shift of the trade-off point for speech component quality and noise attenuation compared to **CEM$_{SI}$**. Both serial approaches manage to consistently mitigate this drawback, where applying **CEM$_{SI}$** first (dotted green line, inverted triangle markers) is able to further improve **CEM$_{SI}$** by up to 0.15 MOS points at an additionally slightly higher noise attenuation. Alternatively, when applying the envelope enhancement first (dashed green line, triangle markers), the **CEM$_{SI}$** baseline can be improved by an average of 0.4 dB NA$_{seg}$, while maintaining a comparable speech component quality.

The data-driven **IRM** baseline shows a surprisingly high speech component quality that is exceeding the performance of all other approaches. However, when further analyzing the

https://doi.org/10.24355/dbbs.084-202102121510-0

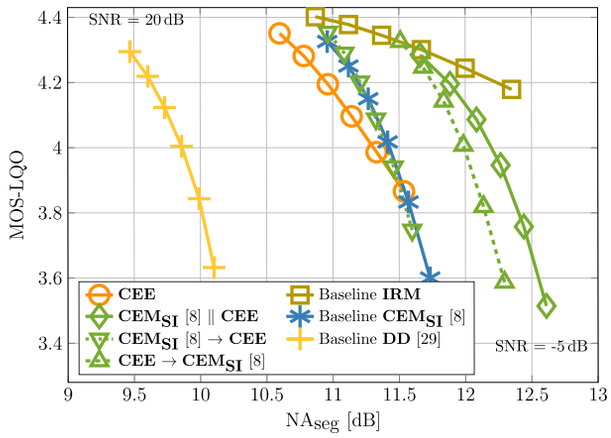


Fig. 12. Comparing the performance of the **CEE** system, the **baselines CEM$_{SI}$**, **IRM**, and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech component quality measured by MOS-LQO and NA$_{seg}$ on the **test set with unseen noise files**.

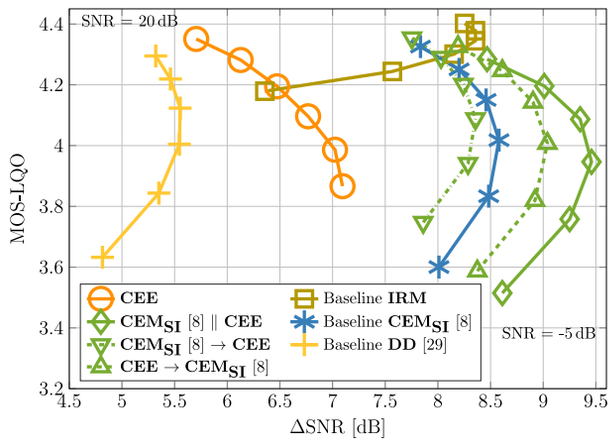


Fig. 13. Comparing the performance of the **CEE** system, the **baselines CEM$_{SI}$**, **IRM**, and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech component quality measured by MOS-LQO and **ΔSNR** on the **test set with unseen noise files**.

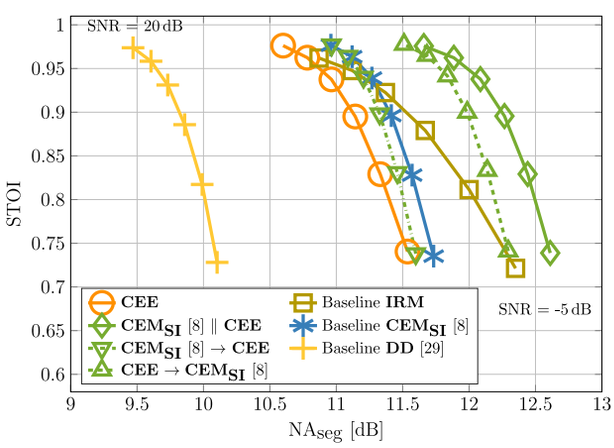


Fig. 14. Comparing the performance of the **CEE** system, the **baselines CEM$_{SI}$**, **IRM**, and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech intelligibility measured by **STOI** and NA$_{seg}$ on the **test set with unseen noise files**.

ΔSNR as shown in Fig. 10, the approach shows the lowest ΔSNR improvement, especially in the important low-SNR conditions. This indicates that the **IRM** approach causes a broadband attenuation of noise *and* speech which is not penalized by PESQ as mentioned in Section V-D. Only in the (not so important) high-SNR conditions the **IRM** approach outperforms the other approaches also in terms of ΔSNR. A further issue with **IRM** is that the residual background noise shows a fluctuating temporal evolvement and thus results in an unsettled subjective listening experience.[2] The **IRM** approach seems to be unable to generate coherent residual background noise which is not surprising, as the neural network has no recurrent modules or any memory which would allow it to produce coherent output w.r.t. previously processed frames. Even though it obtains high NA$_{seg}$ results, the **CEE** approach also shows that the ΔSNR improvement is quite limited. Nonetheless, an improvement over the **DD** baseline, except for the 20 dB condition, is obtained. The proposed serial approach (**CEE** first) takes most profit from the combination of both methods and shows a small but consistent improvement over **CEM$_{SI}$**.

In Fig. 11 we present the intelligibility results measured with STOI for the different approaches. All methods perform similar on STOI, with **IRM** being best in NA$_{seg}$—with the known ΔSNR issue and residual noise quality issue[2] as discussed before.

Furthermore, we have investigated the performance of all the seven depicted approaches on the clean speech data of the test set without noise. Hence, it is not possible to report NA$_{seg}$, but PESQ scores are higher or equal than 4.43 MOS points and STOI is higher or equal than 0.981 for all approaches. This shows that the approaches do not significantly degrade speech quality or intelligibility in clean conditions.

Informal expert listening tests and spectrogram analyses[3] have shown that the parallel and serial (**CEE** first) approaches result in a much smoother and more natural background noise, even in babble noise, owing to the introduced **CEE** method. The approaches also manage to reduce the noise between harmonics facilitated by the integration of the **CEM$_{SI}$** method.

3) Evaluation on the Test Set with Unseen Noise Files: Finally, in Fig. 12 we evaluate the performance on the test set with four unseen noise files, where three are quite non-stationary. The files[4] are taken from the ETSI noise database [44]. Here, the solo **CEE** approach (solid orange line, circle markers) obtains up to 1.4 dB higher NA$_{seg}$, compared to the **DD** baseline and also improves the speech component quality significantly. The performance of the parallel approach (solid green line, diamond markers) is comparable to Fig. 9, where the NA$_{seg}$ is increased at the cost of a lower speech component quality. This is also a general difference between Figs. 9 and 12, since the NA$_{seg}$ in Fig. 12 is consistently lower and thus allows to obtain a higher PESQ score as the classical trade-off. This can be dedicated to

[2]Audio samples for the **IRM** baseline can be found under: https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2018-taslp-cee/

[3]Audio samples can be found under: https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2018-taslp-cee/

[4]`Fullsize_Car1_80Kmh`, `Outside_Traffic_Crossroads`, `Pub_Noise_Binaural_V2`, `Work_Noise_Office_Callcenter`

https://doi.org/10.24355/dbbs.084-202102121510-0

the different noise types, as for Fig. 9 more stationary noise files have been included in the evaluation, which are naturally easier to process than non-stationary noise types, which are predominant in the data for Fig. 12.

The **IRM** baseline shows less improvement w.r.t. NA_{seg} compared to the test set with seen noise files. However, the speech component quality is still quite high, while showing clear detriments in the SNR improvement, as can be seen in Fig. 13. This indicates again that there is also quite some speech attenuation, which is also reflected in STOI (Fig. 14). Here, the **IRM** baseline is outperformed by our serial approach (**CEE** first) and also our parallel approach, where both also show convincing performance in Fig. 13 by improving the SNR consistently.

The serial approach with $\mathbf{CEM_{SI}}$ first (dotted green line, inverted triangle markers) also shows only limited improvement over the $\mathbf{CEM_{SI}}$ baseline (solid blue line, asterisk markers), mainly resulting in an improved speech component quality with a comparable NA_{seg}. However, when applying **CEE** first (dashed green line, triangle markers), we again consistently outperform the $\mathbf{CEM_{SI}}$ baseline by up to more than 0.5 dB NA_{seg}, while obtaining all its benefits even in non-stationary and unseen noise files. Thus, from the various schemes we have proposed in this paper, this is the strongest approach.

VII. Conclusions

We investigated several methods of spectral envelope estimation in the cepstral domain for *a priori* SNR estimation and evaluated their performance in a speech enhancement task with MMSE spectral amplitude estimation. Replacing a hidden Markov model by a deep neural network improves the state accuracy by more than 13% absolute. Evaluated on non-stationary and unseen noise files, the cepstral envelope estimation (CEE) approach alone shows significant improvement over the decision-directed (DD) estimator by up to 1.4 dB noise attenuation (NA), also significantly improving the speech component quality.

The combination with cepstral excitation manipulation (CEM with CEE first) provides a gain of 0.5 dB over CEM and of up to 2 dB over DD in terms of NA, without degrading the speech component quality or intelligibility. The proposed combination also obtains considerable SNR improvement over the baselines in the important low-SNR conditions.

There is still some room for improvement, as shown by the difference in the performance obtained with oracle envelopes and estimated envelopes. Future work will comprise the investigation of how to further reduce this gap, e.g., by more advanced topologies of neural networks which could lead to higher classification accuracies.

References

[1] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 4266–4269.

[2] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[3] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 1036–1039, Dec. 2007.

[4] T. Gerkmann and R. C. Hendriks, "Improved MMSE-based noise psd tracking using temporal cepstrum smoothing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, May 2012, pp. 105–108.

[5] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Mar. 2008, pp. 4897–4900.

[6] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[7] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Two-stage speech enhancement with manipulation of the cepstral excitation," in *Proc. 5th Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, San Francisco, CA, USA, Mar. 2017, pp. 106–110.

[8] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous a priori snr estimation by cepstral excitation manipulation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[9] F. Deng and C. Bao, "Speech enhancement based on ar model parameters estimation," *Speech Commun.*, vol. 79, pp. 30–46, May 2016.

[10] R. Rehr and T. Gerkmann, "A combination of pre-trained approaches and generic methods for an improved speech enhancement," in *Proc. ITG Conf. Speech Commun.*, Paderborn, Germany, Oct. 2016, pp. 51–55.

[11] T. Yoshioka and T. Nakatani, "Speech enhancement based on log spectral envelope model and harmonicity-derived spectral mask, and its coupling with feature compensation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 5064–5067.

[12] S. Srinivasan and J. Samuelsson, "Speech enhancement using a-priori information," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1405–1408.

[13] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[14] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[15] T. Rosenkranz, "Noise codebook adaptation for codebook-based noise reduction," in *Proc. 12th Int. Workshop Acoust. Echo Noise Control*, Tel Aviv, Israel, Aug. 2010.

[16] U. Şimşekli, J. Le Roux, and J. R. Hershey, "Non-negative source-filter dynamical system for speech enhancement," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 6206–6210.

[17] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*. New York, NY, USA: Elsevier Science, 1995, pp. 495–518.

[18] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[19] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 882–892, Mar. 2007.

[20] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, Apr. 2015, pp. 4390–4394.

[21] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[22] S. Mirsamadi and I. Tashev, "Causal speech enhancement combining data-driven learning and suppression rule estimation," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 2870–2874.

[23] J. Abel, M. Strake, and T. Fingscheidt, "Artificial bandwidth extension using deep neural networks for spectral envelope estimation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Xi'an, China, Sep. 2016, pp. 1–5.

[24] J. Abel and T. Fingscheidt, "A DNN regression approach to speech enhancement by artificial bandwidth extension," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Dec. 2017, pp. 219–223.

[25] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 71–83, Jan. 2018.

https://doi.org/10.24355/dbbs.084-202102121510-0

[26] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[27] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[28] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[29] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[30] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, 2005.
[31] B. Fodor and T. Fingscheidt, "MMSE speech enhancement under speech presence uncertainty assuming (generalized) gamma priors throughout," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 4033–4036.
[32] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, USA, May 1996, pp. 629–632.
[33] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, Feb. 1967.
[34] T. Fingscheidt, C. Beaugeant, and S. Suhadi, "Overcoming the statistical independence assumption w.r.t. frequency in speech enhancement," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, USA, Mar. 2005, pp. 1081–1084.
[35] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Upper Saddle River, NJ, USA: Prentice-Hall, 1987.
[36] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.
[37] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
[38] G. McLachlan and D. Peel, *Finite Mixture Models*. Hoboken, NJ, USA: Wiley, 2000.
[39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statis.*, Sardinia, Italy, May 2010, vol. 9, pp. 249–256.
[40] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, 2000.
[41] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
[42] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[43] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010, pp. 3110–3113.
[44] *Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database*, ETSI EG 202 396-1, Sep. 2008.
[45] J. S. Garofolo *et al.*, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.
[46] *Super Wideband Stereo Speech Database*. San Jose, CA, USA: NTT Advanced Technology Corporation.
[47] *Objective Measurement of Active Speech Level*, ITU-T Rec. P.56, Dec. 2011.
[48] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A MATLAB-like environment for machine learning," in *Proc. BigLearn, NIPS Workshop*, Sierra Nevada, Spain, Dec. 2011, pp. 1–6.
[49] S. Gustafsson, R. Martin, and P. Vary, "On the optimization of speech enhancement systems using instrumental measures," in *Proc. Workshop Quality Assessment Speech, Audio, Image Commun.*, Darmstadt, Germany, Mar. 1996, pp. 36–40.
[50] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008.
[51] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, ITU-T Rec. P.862, Feb. 2001.
[52] *Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO*, ITU-T Rec. P.862.1, Nov. 2003.
[53] *Narrow-Band Hands-Free Communication in Motor Vehicles*, ITU-T Rec. P.1100, Jan. 2015.
[54] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

Samy Elshamy received the B.Sc. degree in bioinformatics from Friedrich-Schiller-Universität Jena, Jena, Germany, in 2011 and the M.Sc. degree in computer science from Technische Universität Braunschweig, Braunschweig, Germany, in 2013. He is currently working toward the Ph.D. degree in the field of speech enhancement at the Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany.

Nilesh Madhu received the Dr.-Ing. degree from the Faculty of Electrical Engineering and Information Sciences, Ruhr-Universität Bochum, Bochum, Germany, in 2009. Following this he received a Marie-Curie experienced researcher fellowship for a two-year postdoctoral stay at the KU Leuven, Belgium, where he successfully applied his signal processing knowledge to the field of hearing prostheses and biomedical signal analysis. From 2011 to 2017, he was with NXP Semiconductors, Belgium, where he held the position of a Principal Scientist within the product line Voice and Audio Solutions. During this period, he and his team worked on developing innovative algorithms for audio and speech enhancement for mobile communications devices. Since December 2017, he has been a Professor for audio and speech processing with Ghent University and imec, Belgium. He is passionate about signal processing and is especially interested in signal detection and enhancement for various applications in the fields of healthcare, automation, and communications.

Wouter Tirry received the M.Sc. degree in physics and the Ph.D. degree in solar physics from the University of Leuven, Leuven, Belgium, in 1994 and 1998, respectively. As a Post-doc, he further pursued his research at the National Centre for Atmospheric Research, Boulder, CO, USA. Since 1999, he has been building up expertise in the domain of speech enhancement for mobile devices at Philips and NXP as a Research Engineer and System Architect. He is currently a Senior Principal with the Product Line Voice and Audio Solutions, NXP, Leuven, Belgium, leading the speech technology development activities.

Tim Fingscheidt (S'93–M'98–SM'04) received the Dipl.-Ing. degree in electrical engineering in 1993 and the Ph.D. degree in 1998 from RWTH Aachen University, Aachen, Germany. He further pursued his work on joint speech and channel coding as a Consultant with the Speech Processing Software and Technology Research Department, AT&T Labs, Florham Park, NJ, USA. In 1999, he joined the Signal Processing Department, Siemens AG (COM Mobile Devices), Munich, Germany, and contributed to speech codec standardization in ETSI, 3GPP, and ITU-T. In 2005, he joined Siemens Corporate Technology, Munich, Germany, leading the speech technology development activities in recognition, synthesis, and speaker verification. Since 2006, he has been a Full Professor with the Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany. His research interests include speech and audio signal processing, enhancement, transmission, recognition, and instrumental quality measures. He received several awards, among them are a prize of the Vodafone Mobile Communications Foundation in 1999 and the 2002 prize of the Information Technology branch of the Association of German Electrical Engineers (VDE ITG). In 2017, he co-authored the ITG award-winning publication, but the ITG prize is only awarded once in a life time. He has been a speaker of the Speech Acoustics Committee ITG AT3 since 2015. From 2008 to 2010, he was an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and since 2011 he serves as a member of the IEEE Speech and Language Processing Technical Committee.

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

Publication V

S. Elshamy, T. Fingscheidt, N. Madhu, and W. Tirry, "A Priori SNR Computation for Speech Enhancement Based on Cepstral Envelope Estimation," in *Proc. of IWAENC*, Tokyo, Japan, Sep. 2018, pp. 531–535

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

A PRIORI SNR COMPUTATION FOR SPEECH ENHANCEMENT BASED ON CEPSTRAL ENVELOPE ESTIMATION

Samy Elshamy, Nilesh Madhu†, Wouter Tirry° and Tim Fingscheidt**

*Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, D–38106 Braunschweig, Germany
†Internet Technology and Data Science Lab, Universiteit Gent - imec, 9052 Gent, Belgium
°NXP Software, Interleuvenlaan 80, B–3001 Leuven, Belgium
{s.elshamy,t.fingscheidt}@tu-bs.de, nilesh.madhu@ugent.be, wouter.tirry@nxp.com

ABSTRACT

In this contribution we present our latest investigations and analysis on a novel *a priori* SNR estimator for speech enhancement applications. It is based on a clean spectral envelope estimation with a deep neural network (DNN) in the cepstral domain. Furthermore, by integrating our cepstral excitation manipulation (CEM) approach into this framework, we obtain not only a smooth and natural background noise experience, but also achieve noise reduction between harmonics which is not possible with low-order models. We investigate the performance of the proposed approach in conjunction with three different spectral weighting rules and show improvement of more than 3.5 dB noise attenuation vs. the well-known decision-directed (DD) approach without a significant trade-off in speech distortion.

Index Terms— a priori SNR, speech enhancement, cepstrum

1. INTRODUCTION

The broad field of speech enhancement comprises various applications that aim to facilitate the communication between human beings. Among them we find speech presence probability estimation, voice activity detection, and, e.g., noise reduction. The latter often uses a real-valued spectral weighting rule [1] in the frequency domain for a bin-wise noise suppression of a noisy microphone signal's amplitudes. These weighting rules are usually a function of the *a priori* signal-to-noise ratio (SNR) and oftentimes also of the *a posteriori* SNR.

The well-known decision-directed (DD) approach [2] defines an *a priori* SNR estimate that depends both on the past *a priori* SNR and the *a posteriori* SNR to obtain the estimate. Although the DD technique suffers from its incapability to track sudden changes of the true SNR, it is still regarded as classical state of the art.

Among the numerous more recent publications that investigate different *a priori* SNR estimation approaches [3, 4, 5, 6, 7], a generalized version of the DD approach has been proposed recently by Chinaev and Haeb-Umbach [8]. The method operates in a generalized spectral domain instead of the power domain. The authors show improved performance for high global SNR conditions for the generalized approach, while the original method, operating in the power domain, shows optimal behavior in low-SNR conditions.

Stahl and Mowlaee introduced a harmonic signal model for *a priori* SNR estimation in [9]. The model allows to interpolate between frequency bins and thus to smooth the *a priori* SNR according to harmonic trajectories. Thereby, the authors show improved noise attenuation capability without introducing additional speech distortion compared to the DD approach.

Furthermore, the incorporation of other models has been investigated in the recent past [10, 11, 12, 13], showing some improvement over the DD approach.

Very recently we proposed a novel *a priori* SNR estimator [12] based on cepstral excitation manipulation (CEM), which exploits the human speech production model. Its core features are the improvement of noise attenuation between harmonics and also the preservation of weak harmonic structures. Therein, we could show a more balanced and thus enhanced performance over the DD approach and also over two further, more recent *a priori* SNR estimators [4, 5]. Accordingly, both the DD and the CEM *a priori* SNR estimator serve as baselines for *this* work. Most recently we proposed a cepstral envelope estimation (CEE) approach [13] that nicely complements the CEM approach by not only enhancing the excitation signal but also the envelope. We described in detail how the proposed envelope estimator has been distilled from various investigated approaches.

In this paper we briefly revisit our findings from [13] and investigate the performance of the CEE approach for *a priori* SNR estimation alone, and in conjunction with CEM. We evaluate the estimators in a speech enhancement task together with three different weighting rules.

This contribution is structured as follows: In Section 2 we briefly introduce the signal model along with the CEE technique and provide insight into our investigations on the different methods. This is followed by a short introduction of the speech enhancement framework and the three weighting rules in Section 3. Subsequently, the experimental setup and the evaluation of our results is presented in Section 4. We finally conclude the paper in Section 5.

2. CEPSTRAL ENVELOPE ESTIMATION (CEE)

The microphone signal $y(n)$ is modeled as the superposition of the time-domain speech signal $s(n)$ and the noise signal $d(n)$ as $y(n) = s(n) + d(n)$, with n as discrete-time sample index. The frequency-domain entities are obtained by applying a K-point discrete Fourier transform as $Y_\ell(k) = S_\ell(k) + D_\ell(k)$, where ℓ represents the frame index and $0 \leq k \leq K-1$ the frequency bin index. Furthermore, we assume that both signals, noise and speech, have zero mean and that they are statistically independent.

The basic idea of our approach (see Figure 1) is to split a preliminary denoised microphone signal $\tilde{Y}_\ell(k)$ into its envelope (Figure 1,

978-1-5386-8151-0/18/$31.00

https://doi.org/10.24355/dbbs.084-202102121510-0

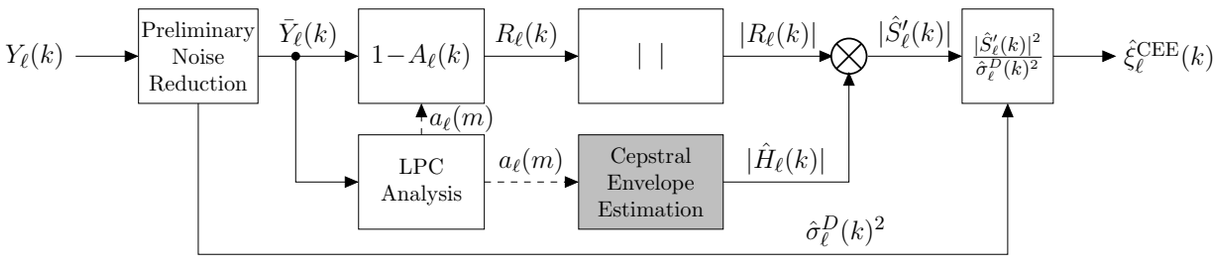


Fig. 1. Block diagram of the **proposed *a priori* SNR estimator** based on **cepstral envelope estimation (CEE)**.

LPC analysis, lower path) and its excitation $R_\ell(k)$ by LPC analysis. The denoised envelope is subsequently replaced by a clean envelope estimate $|\hat{H}_\ell(k)|$ and mixed with the excitation signal. It is used further with the noise power estimate $\hat{\sigma}_\ell^D(k)^2$ from the preliminary noise reduction to calculate the *a priori* SNR $\hat{\xi}_\ell^{\mathrm{CEE}}(k)$. The estimation is done in the cepstral domain by converting (Figure 2, feature conversion block) the $N = 10$ LPC coefficients to $N + 1 = 11$ cepstral coefficients using [14] as

$$c_\ell^H(m) = a_\ell(m) + \frac{1}{m}\sum_{\mu=1}^{m-1}\Big[(m-\mu)\cdot a_\ell(\mu)\cdot c_\ell^H(m-\mu)\Big] \quad (1)$$

for $1 \le m \le N$ and

$$c_\ell^H(m=0) = 0 = \log(P_p = 1) \quad (2)$$

for $m = 0$. We set the prediction error power P_p to a fixed value to obtain envelopes with a comparable energy level. This allows us to work with N coefficients only, as the first coefficient has the same value (zero) for all vectors. After estimating the clean envelope representing coefficients $c_\ell^{\hat{H}}(m)$ (see Figure 2, bold face for vector notation) we convert them back to LPC coefficients with [14]

$$\hat{a}_\ell(m) = c_\ell^{\hat{H}}(m) - \frac{1}{m}\sum_{\mu=0}^{m-1}\Big[(m-\mu)\cdot c_\ell^{\hat{H}}(m-\mu)\cdot \hat{a}_\ell(\mu)\Big] \quad (3)$$

for $1 \le m \le N$. The spectral representation $|\hat{H}_\ell(k)|$ is obtained from

$$|\hat{H}_\ell(k)| = \frac{1}{|1-\hat{A}_\ell(k)|}, \quad (4)$$

where $\hat{A}_\ell(k)$ is calculated by applying a K-point DFT to the zero-padded LPC coefficients $\hat{a}_\ell(m)$. This is done in the spectral conversion block in Figure 2.

We have evaluated and optimized different approaches for the cepstral envelope estimation task in [13]. We started with a classic hidden Markov model (HMM) with Gaussian mixture models (GMMs) as acoustic backend (GMM-HMM) where the hidden states represent clean, and the observations denoised envelopes. We found out by means of an oracle experiment that a codebook size of $64 + 1 = 65$ is sufficient. The codebook entries are obtained by using the Linde-Buzo-Gray [15] algorithm and the extra entry is exclusively representing non-speech envelopes. Best posterior state probability accuracy was obtained by choosing 16 modes for the GMMs. Furthermore, we investigated maximum a posteriori (MAP) and also minimum mean-square error (MMSE) estimation, resulting in superior performance of the latter in our noise reduction task.

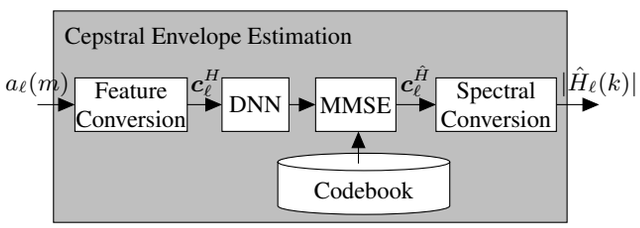


Fig. 2. Block diagram of the **preferred cepstral envelope estimation (CEE) method** using a classification DNN together with MMSE estimation.

Hence, we fixed the number of parameters and investigated the replacement of the GMMs by a classification deep neural network (DNN). We trained differently configured networks with up to six hidden layers, making sure that the aberration of parameters is always less than 2% by adjusting the number of nodes per layer, accordingly. Based on the best state posterior probability accuracy on speech active frames we found a network with six hidden layers, 58 nodes per layer, and sigmoid activation function to be optimal for classification. Thereby, the overall accuracy could be increased by 10 % absolute on the development set compared to the GMM-HMM approach. However, the incorporation of the DNN into the HMM yielded only comparable performance of the DNN-HMM vs. the GMM-HMM. Subsequently, we replaced the whole HMM structure by the classification DNN and could further improve the performance, now being able to fully benefit from the additional 10 % accuracy on the development set.

In Figure 2 we depict the processing structure of our favored estimator and refer to this method as CEE throughout the remainder of this paper. We have also investigated the performance of a DNN trained in regression mode, to directly estimate clean envelope-representing coefficients from the denoised observation. An optimal configuration was found for six hidden layers, 58 nodes, and also sigmoid activation function, but the performance in our noise reduction task was imbalanced, showing some detriments in the low-SNR conditions. So our proposal is to use the aforementioned classification DNN with subsequent codebook-supported MMSE estimation, as shown in Figure 2.

3. SPEECH ENHANCEMENT FRAMEWORK

Later evaluation of the *a priori* SNR estimators will be conducted in a speech enhancement framework consisting of a minimum statistics

https://doi.org/10.24355/dbbs.084-202102121510-0

noise power estimator [16], the *a priori* SNR estimator under test, and a spectral weighting rule to obtain the enhanced speech signal as

$$\hat{S}_\ell(k) = G_\ell(k) \cdot Y_\ell(k). \tag{5}$$

The spectral weighting rules $G_\ell(k) = f(\hat{\xi}_\ell(k), \gamma_\ell(k))$ are the minimum mean square error log-spectral amplitude (MMSE-LSA) estimator [17], the Wiener filter (WF) [18], and the super-Gaussian joint maximum a posteriori (SG-jMAP) estimator [19]. An *a posteriori* SNR

$$\gamma_\ell(k) = \frac{|Y_\ell(k)|^2}{\sigma_\ell(k)^2} \tag{6}$$

is required for the MMSE-LSA and the SG-jMAP spectral weighting rule, and also for the DD *a priori* SNR baseline estimator according to [2]

$$\hat{\xi}_\ell^{\mathrm{DD}}(k) = \\ (1-\beta_{\mathrm{DD}}) \cdot \max\{\hat{\gamma}_\ell(k)-1, 0\} + \beta_{\mathrm{DD}} \frac{|\hat{S}_{\ell-1}(k)|^2}{\hat{\sigma}_{\ell-1}^D(k)^2}. \tag{7}$$

The $\mathrm{CEM_{SI}}$ baseline [12] is refining a clean speech amplitude estimate in an instantaneous fashion by modifying the excitation signal based on pre-trained templates. Is is subsequently used with the noise power estimate from the preliminary noise reduction to obtain $\hat{\xi}_\ell^{\mathrm{CEM}}(k)$.

If our proposed CEE-based *a priori* SNR estimator according to Figure 2 is employed, the minimum statistics noise power estimator is executed as part of the preliminary noise reduction, which internally also contains a DD *a priori* SNR estimation and an MMSE-LSA weighting rule. The rest of our CEE *a priori* SNR estimator is shown in Figure 1 with Figure 2 as discussed.

We will also investigate an approach that concatenates the CEE *a priori* SNR estimator with the CEM technique from [12]. For that purpose the preliminary noise reduction as it is required for the CEM approach consists of the complete Figure 1, including a subsequent MMSE-LSA spectral weighting rule which is applied to the microphone signal. The further processing according to [12] provides then the final *a priori* SNR estimate $\hat{\xi}_\ell^{\mathrm{CEE}\to\mathrm{CEM}}(k)$ — for details please be referred to [12, 13]. Note that in this serial approach the noise power estimate that is used throughout is the one of the aforementioned preliminary noise reduction in the CEE approach, see Figure 1.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

The DD estimator, wherever it is employed, is tuned with optimal parameters[1] [20] for each weighting rule. The DD estimator as part of the preliminary noise reduction in Figure 1 uses parameters as shown[1] for MMSE-LSA, since this is the weighting rule of the preliminary noise reduction. We work with a sample rate of 8 kHz, a frame size of $K = 256$ samples with a frame shift of 50 %. For analysis and overlap-add synthesis we utilize a periodic square root Hann window. The training and development sets for the investigations in Section 2 are taken from the TIMIT database [21]. The clean speech is mixed at six different SNR conditions ranging from -5 dB to 20 dB in 5 dB steps together with disjoint portions of 53 noise files taken from the ETSI [22] and the QUT [23] noise databases. For a test set with unseen noise files we use four files[2] from the ETSI database exclusively which are not used for training or development. However, similar noise types have been used also for the training process. Signal levels are adjusted according to ITU-T P.56 [24] and subsequently superimposed.

[1] Optimal parameters for the DD estimator and each weighting rule:

MMSE-LSA: $\beta_{\mathrm{DD}} = 0.975, \xi_{\min} = -15$ dB
SG-jMAP: $\beta_{\mathrm{DD}} = 0.993, \xi_{\min} = -14$ dB
WF: $\beta_{\mathrm{DD}} = 0.99, \xi_{\min} = -14$ dB.

4.2. Quality Measures

To evaluate the estimators in a speech enhancement task we use the white-box approach [25] which allows us to evaluate the *filtered* speech component $\tilde{s}(n)$ and the *filtered* noise component $\tilde{d}(n)$ of the enhanced signal $\hat{s}(n)$, separately. This is done by applying the final gain function $G_\ell(k)$ not only to the microphone signal $Y_\ell(k)$ in order to obtain the enhanced speech $\hat{S}_\ell(k)$, but also to the separate speech component $S_\ell(k)$ and noise component $D_\ell(k)$, followed by inverse DFT and overlap-add synthesis. As objective measures we use the segmental speech-to-speech-distortion ratio (SSDR) [26] which is calculated as

$$\mathrm{SSDR_{seg}} = \frac{1}{|\mathcal{L}_1|} \sum_{\ell \in \mathcal{L}_1} \mathrm{SSDR}(\ell) \tag{8}$$

with $\mathcal{L}_1$ being the set of speech active frames,

$$\mathrm{SSDR}(\ell) = \max\left\{\min\left\{\mathrm{SSDR}'(\ell), R_{\max}\right\}, R_{\min}\right\} \tag{9}$$

where $R_{\max}$ and $R_{\min}$ limit the values to 30 dB and -10 dB, respectively. The frame-wise ratio is obtained as

$$\mathrm{SSDR}'(\ell) = 10\log_{10}\left[\frac{\sum_{\nu=0}^{N-1} s(\nu+\ell N)^2}{\sum_{\nu=0}^{N-1} e(\nu+\ell N)^2}\right] \tag{10}$$

with the error signal being

$$e(\nu+\ell N) = \tilde{s}(\nu+\ell N+\Delta) - s(\nu+\ell N). \tag{11}$$

The term Δ is accounting for potential processing delay and ℓ is depicting a segment of length $N = 256$ samples. A high $\mathrm{SSDR_{seg}}$ indicates a strong similarity of the speech component with respect to the clean reference signal.

To account for the noise attenuation we additionally report the ΔSNR which is a global measure and calculated as

$$\Delta\mathrm{SNR} = \mathrm{SNR_{out}} - \mathrm{SNR_{in}}, \tag{12}$$

where $\mathrm{SNR_{out}}$ is the SNR of the *filtered* components $\tilde{s}(n)$ and $\tilde{d}(n)$, and $\mathrm{SNR_{in}}$ is the SNR of the unprocessed components $s(n)$ and $d(n)$. Both SNRs are measured in line with ITU-T P.56 [24] where for the speech signals only speech active portions are considered. The ΔSNR gives information on the global SNR improvement by considering both components simultaneously.

4.3. Discussion

In Figure 3 we depict the results for the different *a priori* SNR estimators under test with the three weighting rules MMSE-LSA, SG-jMAP, and WF. We plot the $\mathrm{SSDR_{seg}}$ vs. the ΔSNR and each marker represents one SNR condition, where -5 dB is at the bottom and

[2] `Fullsize_Car1_80Kmh`, `Outside_Traffic_Crossroads`, `Pub_Noise_Binaural_V2`, `Work_Noise_Office_Callcenter`

https://doi.org/10.24355/dbbs.084-202102121510-0

20 dB is at the top in steps of 5 dB. In general, the WF seems to achieve the highest ΔSNR for each approach, while the speech component quality suffers, which is quite obvious especially for the **DD** approach. The most recent weighting rule SG-jMAP provides best speech component quality among the analyzed estimators, however, offering less noise attenuation as a typical trade-off. The MMSE-LSA estimator settles somewhere in between showing a balanced performance of the *a priori* SNR estimators.

The **CEE** *a priori* SNR estimator (solid orange line, asterisk markers) outperforms the **DD** baseline (solid yellow line, plus markers) by about 2 dB ΔSNR for the MMSE-LSA and SG-jMAP weighting rules in the -5 dB SNR condition. Using the SG-jMAP weighting rule, **CEE** exceeds the performance of the **DD** approach also consistently in terms of SSDR_{seg}. When used with the WF, only the important low-SNR conditions show reasonable performance gain for **CEE**.

The recently published $\textbf{CEM}_{\textbf{SI}}$ baseline [12] (solid green line, square markers) exceeds clearly the **DD** baseline, and also **CEE** when operating alone, in terms of noise attenuation for every weighting rule owing to its ability to effectively reduce noise between the harmonics. The highest performance gain obtained over **DD** amounts to more than 3 dB ΔSNR for $\textbf{CEM}_{\textbf{SI}}$ when either using MMSE-LSA or SG-jMAP. This gain can be further enlarged by concatenating (symbol →) the **CEE** approach with the $\textbf{CEM}_{\textbf{SI}}$ baseline (dashed green line, triangle markers). Thereby, we obtain a ΔSNR that is higher by more than 3.5 dB compared to the DD approach for the MMSE-LSA weighting rule.

The investigated approaches (**CEE** and $\textbf{CEE}\rightarrow\textbf{CEM}_{\textbf{SI}}$) appear to be more robust compared to **DD** as the speech component quality remains comparable for the respective approach when exchanging MMSE-LSA by the WF, while simultaneously also showing higher ΔSNR. Here, the **DD** approach experiences quite some negative effects on the speech component quality due to the increase of noise attenuation. Hence, we recommend the serial approach $\textbf{CEE}\rightarrow\textbf{CEM}_{\textbf{SI}}$ as it offers robustness across various weighting rules while being able to mitigate the classical trade-off between speech component quality and noise attenuation. Informal expert analysis and listening tests[3] have shown that the approach results in a very smooth and also natural sound of the remaining low-level background noise. This is an advantage over both baselines, **DD** and also $\textbf{CEM}_{\textbf{SI}}$.

5. CONCLUSIONS

We investigated the performance of a novel *a priori* SNR estimator in a noise reduction environment with three different spectral weighting rules. We could show that the proposed serial estimator, which uses cepstral envelope estimation (CEE) in conjunction with cepstral excitation manipulation (CEM), exceeds CEM consistently by up to 0.4 dB ΔSNR, even in non-stationary noise, and improves by more than 3.5 dB vs. the decision-directed (DD) approach. At the same time, no significant trade-off in speech distortion is observed.

6. REFERENCES

[1] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*, Springer, Berlin, 2008.

[3] Audio samples can be found under:
https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2018-iwaenc-cee/

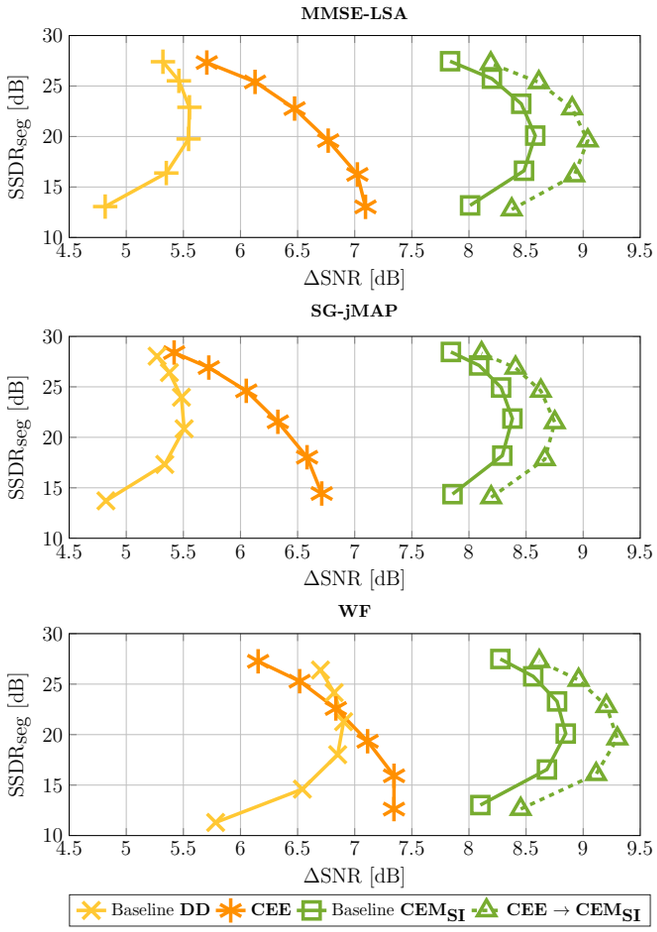


Fig. 3. Evaluation of SSDR_{seg} and ΔSNR for the *a priori* SNR estimators under test in non-stationary and unseen noises, together with three different spectral weighting rules.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[3] I. Cohen, "Relaxed Statistical Model for Speech Enhancement and A Priori SNR Estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, Sep 2005.

[4] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[5] C. Breithaupt, T. Gerkmann, and R. Martin, "A Novel A Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing," in *Proc. of ICASSP*, Las Vegas, NV, USA, Mar. 2008, pp. 4897–4900.

[6] S. Suhadi, C. Last, and T. Fingscheidt, "A Data-Driven Approach to A Priori SNR Estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[7] H. S. Shin, T. Fingscheidt, and H.-G. Kang, "A Priori SNR Estimation Using Air and Bone-Conduction Microphones,"

https://doi.org/10.24355/dbbs.084-202102121510-0

IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 2015–2025, Nov. 2015.

[8] A. Chinaev and R. Haeb-Umbach, "A Priori SNR Estimation Using a Generalized Decision Directed Approach," in *Proc. of Interspeech*, San Francisco, CA, USA, Sept. 2016, pp. 3758–3762.

[9] J. Stahl and P. Mowlaee, "A Simple and Effective Framework for A Priori SNR Estimation," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 5644–5648.

[10] S. Elshamy, N. Madhu, W. J. Tirry, and T. Fingscheidt, "An Iterative Speech Model-Based A Priori SNR Estimator," in *Proc. of Interspeech*, Dresden, Germany, Sept. 2015, pp. 1740–1744.

[11] A. Chinaev, J. Heitkaemper, and R. Haeb-Umbach, "A Priori SNR Estimation Using Weibull Mixture Model," in *Proc. of ITG Conference on Speech Communication*, Paderborn, Germany, Oct. 2016, pp. 297–301.

[12] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous A Priori SNR Estimation by Cepstral Excitation Manipulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[13] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "DNN-Supported Speech Enhancement With Cepstral Estimation of Both Excitation and Envelope," *Submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[14] P. E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice Hall, Inc., Upper Saddle River, NJ, USA, 1987.

[15] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," in *IEEE Transactions on Communications*, Jan. 1980, vol. 28, pp. 84–95.

[16] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[17] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[18] P. Scalart and J. V. Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP*, Atlanta, GA, USA, May 1996, pp. 629–632.

[19] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[20] H. Yu, *Post-Filter Optimization for Multichannel Automotive Speech Enhancement*, Ph.D. thesis, Institute for Communications Technology, Technische Universität Braunschweig, 2013.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium (LDC), 1993.

[22] ETSI, *EG 202 396-1: Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database*, European Telecommunications Standards Institute, Sept. 2008.

[23] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT Corpus for the Evaluation of Voice Activity Detection Algorithms," in *Proc. of Interspeech*, Makuhari, Japan, Sept. 2010, pp. 3110–3113.

[24] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.

[25] S. Gustafsson, R. Martin, and P. Vary, "On the Optimization of Speech Enhancement Systems Using Instrumental Measures," in *Proc. of Workshop on Quality Assessment in Speech, Audio, and Image Communication*, Darmstadt, Germany, Mar. 1996, pp. 36–40.

[26] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-Optimized Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, May 2008.

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

Publication VI

S. Elshamy and T. Fingscheidt, "DNN-Based Cepstral Excitation Manipulation for Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1803–1814, Nov. 2019

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

DNN-Based Cepstral Excitation Manipulation for Speech Enhancement

Samy Elshamy and Tim Fingscheidt, *Senior Member, IEEE*

***Abstract*—This contribution aims at speech model-based speech enhancement by exploiting the source-filter model of human speech production. The proposed method enhances the excitation signal in the cepstral domain by making use of a deep neural network (DNN). We investigate two types of target representations along with the significant effects of their normalization. The new approach exceeds the performance of a formerly introduced classical signal processing-based cepstral excitation manipulation (CEM) method in terms of noise attenuation by about 1.5 dB. We show that this gain also holds true when comparing serial combinations of envelope and excitation enhancement. In the important low-SNR conditions, no significant trade-off for speech component quality or speech intelligibility is induced, while allowing for substantially higher noise attenuation. In total, a traditional purely statistical state-of-the-art speech enhancement system is outperformed by more than 3 dB noise attenuation.**

***Index Terms*—Speech enhancement, deep learning, cepstrum, *a priori* SNR.**

I. Introduction

SPEECH enhancement is still a very important and active field of research. Its primary aim is to improve speech quality and intelligibility, to facilitate the most natural way of communication. Speech signals might be corrupted by, e.g., bandwidth limitation, coupling of noise, echo, and reverberation. In order to combat such problems, various algorithms have been developed and improved over time.

Single-channel noise reduction is still a challenging task, which is addressed here. Even though traditional systems might be still considered as state of the art, recent advances in speech enhancement make more and more use of modern deep learning technologies and often end-to-end solutions are presented (e.g., [1]–[3]). As mentioned in [3], one issue of conventional DNN-based enhancement models is the discontinuity of the enhanced signals when processed in a frame-based manner. The authors resolve the problem by enhancing whole utterances on waveform level which requires the availability of complete recordings or at least a very large buffer. This is not applicable for telephony applications, where delay has to be as low as possible and frame-wise processing is essential. In the following, more recent advances will be presented briefly.

Manuscript received April 9, 2019; revised July 3, 2019 and August 2, 2019; accepted August 2, 2019. Date of publication August 8, 2019; date of current version August 21, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. S. Doclo. *(Corresponding author: Samy Elshamy.)*

The authors are with the Institute for Communications Technology, Technische Universität Braunschweig, 38106 Braunschweig, Germany (e-mail: s.elshamy@tu-bs.de; t.fingscheidt@tu-bs.de).

Digital Object Identifier 10.1109/TASLP.2019.2933698

A sketch of less holistic approaches, that in parts still respect traditional and statistical speech enhancement is shown in [4]. The publication nicely shows various levels of granularity that allow to move away from end-to-end solutions towards more traditional structures, still being able to benefit from modern technology. Following this, DNN-based learning of spectral weighting rules has been evaluated, e.g., for ideal binary masks and ideal ratio masks in [5], [6].

The spectral envelope codebook-based work by Srinivasan *et al.* [9]–[11] was brought from an autoregressive (AR) model to the cepstral domain by Rosenkranz *et al.* [12], and it has been picked up again recently in [13] and [14]. In [13], the authors combine the existing auto regressive-based approach with a noise estimator [15] to circumvent the dependency on a noise codebook. Additionally, they introduce an SPP estimator [16] to combat the lack of noise suppression between the harmonics, which is naturally not possible when only spectral envelopes are used for the estimation of the clean speech. This issue has been further addressed in our previous work [7] and is also investigated together with the preservation of harmonics in this publication by analyzing the effects of the normalization of targets during the training process. The authors in [14] aim to replace both codebooks by estimating the parameters of the AR models for speech and noise simultaneously with a single network that predicts line spectral pairs. In order to combat the inability to reduce noise between the harmonics, they also use the SPP estimator from [16]. In both cases the estimated entities are used to create a Wiener filter and for the latter approach it depicts a further step towards a more modular integration of DNNs into a statistical speech enhancement framework.

The *a priori* SNR represents a more generic entity, as it can be easily plugged into various statistical systems, also being a key factor in noise reduction. It has been subject to research not only through the past decades [17]–[19], but particularly in the recent past with quite some success [7], [8], [20]–[25]. While most approaches work in the frequency domain, Breithaupt *et al.* originally pioneered the way for *a priori* SNR estimation in the cepstral domain [20]. Stahl *et al.* pick up the original decision-directed (DD) approach by Ephraim and Malah [17] and propose to smooth the *a priori* SNR not over isolated frequency bins but with respect to harmonic trajectories [24]. This leads to higher noise attenuation without further speech distortion. Xu *et al.* make use of discriminative non-negative matrix factorization (DNMF) for *a priori* SNR estimation and present two different approaches [25]. One approach uses DNMF to estimate speech and noise power to directly calculate the *a priori* SNR, while

https://doi.org/10.24355/dbbs.084-202102121510-0

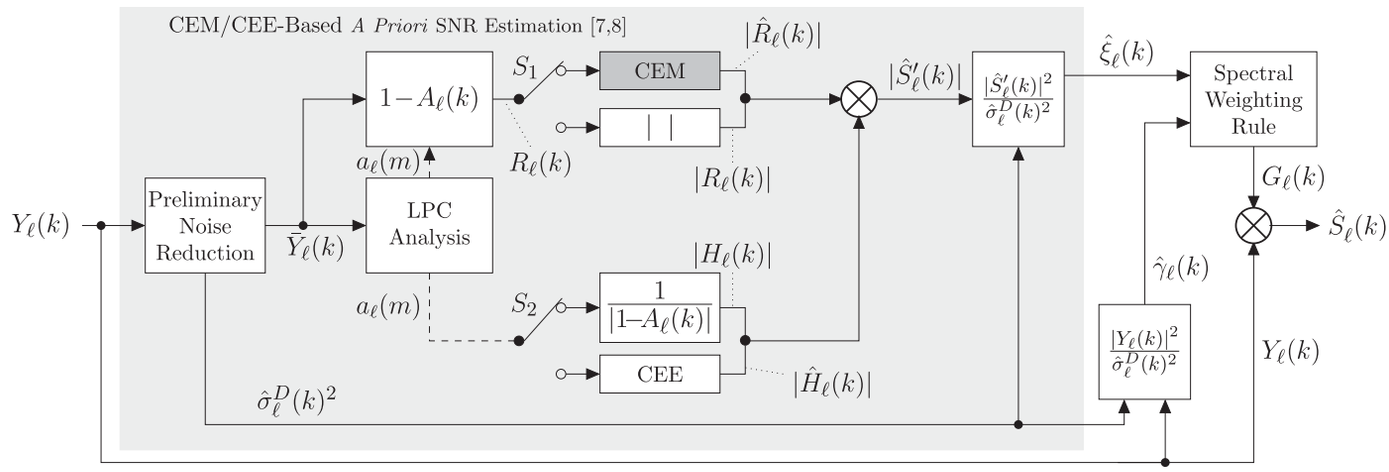


Fig. 1. Schematic of the speech enhancement framework with either **cepstral excitation manipulation (CEM)** [7] (switches S_1 and S_2 as shown) or **cepstral envelope estimation (CEE)** [8] (switches S_1 and S_2 in lower position) *a priori* SNR estimation. The CEM block is depicted in more detail in Fig. 2, as its replacement by a deep neural network is core novelty of this work.

the other uses DNMF to only estimate the noise power which is then used together with the DD approach. Both methods obtain better results than DNMF approaches that are commonly used to directly estimate the clean speech. However, they rely on noise codebooks which might limit the capability of generalization.

In this contribution, we aim to exploit the potential of the cepstral excitation manipulation (CEM) approach further, as the current state-of-the-art CEM solution [7] offers room for improvement, in terms of speech quality, speech intelligibility, and also noise attenuation. To do so, we incorporate deep neural network (DNN) models to enhance the residual signal for the purpose of *a priori* SNR estimation for speech enhancement. A particular aspect is that the explicit F_0 estimator as required by state-of-the-art CEM is not needed anymore for the core functionality of CEM in our new approach. We investigate two different lines of research for the *a priori* SNR numerator. The first aims to restore the clean speech residual signal from a noisy observation. The second is to restore the clean speech signal itself by estimating a residual signal which is also considering and compensating the degeneration of the spectral envelope in noisy conditions. The performance of the *a priori* SNR estimator is evaluated in a speech enhancement task—although its application is not limited to that—and measured by renowned metrics such as the PESQ score [26], [27], the short-time objective intelligibility measure (STOI) [28], and also the segmental noise attenuation (NA_{seg}) [29].

This paper is structured as follows. We briefly describe the signal model and speech enhancement framework in Section II, followed by the introduction of the baseline approaches in Section III. Next, we present the new DNN-based CEM approach in Section IV, and subsequently depict our experimental setup in Section V. Finally, we evaluate, discuss, and conclude the paper in Section VI and Section VII, respectively.

II. Signal Model and Speech Enhancement Framework

In this section we briefly introduce our signal model and the speech enhancement framework which is used for some preliminary experiments and for the evaluation.

A. Signal Model

We model the noisy time-domain microphone observation as

$$y(n) = s(n) + d(n), \tag{1}$$

where $s(n)$ is the clean speech component, $d(n)$ the noise component, and n the discrete-time sample index. Both components are superimposed to obtain the microphone signal $y(n)$. We apply a K-point discrete Fourier transform (DFT) to obtain the corresponding frequency domain representation as

$$Y_\ell(k) = S_\ell(k) + D_\ell(k), \tag{2}$$

with frequency bin index $0 \leq k \leq K-1$ and frame index ℓ. Furthermore, we assume zero-mean speech and noise signals.

B. Speech Enhancement Framework

The speech enhancement framework we are utilizing is depicted in Fig. 1. It is starting with a preliminary noise reduction which is intended to process the noisy microphone signal $Y_\ell(k)$ in a first stage to provide a more suitable input signal $\bar{Y}_\ell(k)$ for the following processing. This first noise reduction stage is not restricted to any specific configuration, however, one should assure matched conditions with any potential training algorithms that might be required for subsequent processing stages. We use the minimum statistics (MS) [30] noise power estimator together with decision-directed (DD) [17] *a priori* SNR estimation and as spectral weighting rule the minimum mean squared error log-spectral amplitude estimator (MMSE-LSA) [31]. This stage is followed by a linear predictive coding (LPC) analysis block which subsequently allows for separate enhancement of the excitation signal $R_\ell(k)$ (upper path) and of the spectral envelope $H_\ell(k)$ (lower path). Both enhancement methods are explained further in more detail in Sections III-A and III-B, respectively. The enhanced signals' spectral amplitudes ($|\hat{R}_\ell(k)|$ or $|\hat{H}_\ell(k)|$) are then mixed with the respective counterpart ($|H_\ell(k)|$ or $|R_\ell(k)|$), to obtain an intermediate clean speech spectral amplitude estimate $|\hat{S}'_\ell(k)|$. It is important to note that—along with the noise power estimate $\hat{\sigma}_\ell^D(k)^2$ from the preliminary noise reduction—this estimate is only used as the numerator for the *a*

https://doi.org/10.24355/dbbs.084-202102121510-0

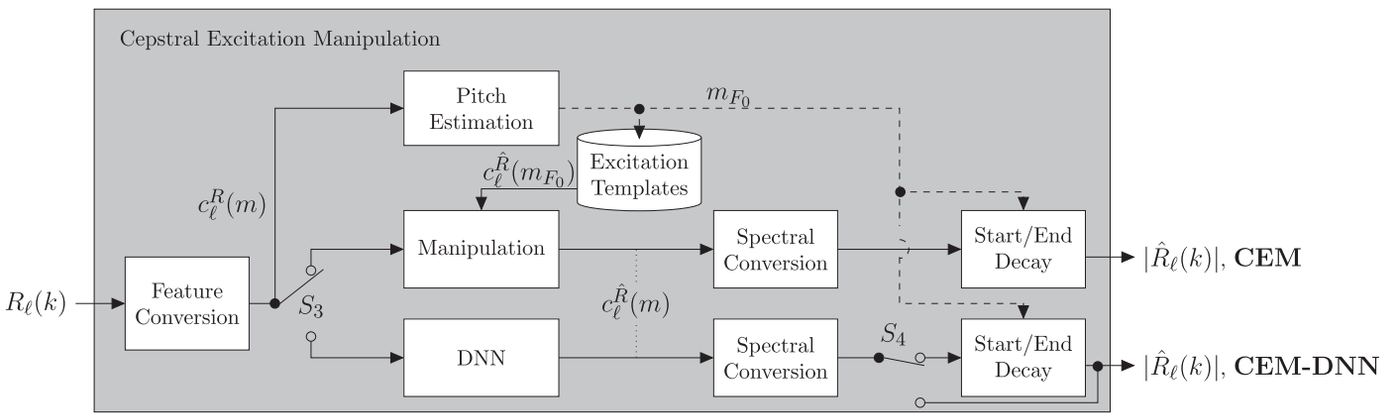


Fig. 2. Block diagram of the **CEM baseline approach** [7] and the **new proposed CEM-DNN approach** which is using a deep neural network (DNN). Here, switch S_3 determines the used algorithm. The **CEM-DNN** method is investigated with and without applied start/end decay which is determined by the position of switch S_4.

priori SNR estimate as

$$\hat{\xi}_\ell(k) = \frac{|\hat{S}'_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}. \tag{3}$$

It is then used jointly with the *a posteriori* SNR estimate $\hat{\gamma}_\ell(k)$ to calculate a spectral weighting rule

$$G_\ell(k) = f(\hat{\xi}_\ell(k), \hat{\gamma}_\ell(k)), \tag{4}$$

which is in our case again the MMSE-LSA estimator [31] for all traditional statistical-based approaches. Finally, the clean speech estimate $\hat{S}_\ell(k)$ is obtained by multiplying the real-valued gain function $G_\ell(k)$, which is limited to $G_{\min} = -15$ dB, with the microphone signal $Y_\ell(k)$ as

$$\hat{S}_\ell(k) = Y_\ell(k) \cdot G_\ell(k). \tag{5}$$

III. Baseline Approaches

As the proposed method builds upon the originally published CEM approach [7], we briefly revisit CEM as it has already shown to improve over common speech enhancement approaches. Among them are traditional statistical-based systems using e.g., the decision-directed *a priori* SNR estimator by Ephraim and Malah [17], the harmonic regeneration approach by Plapous *et al.* [19], and also the selective cepstro-temporal smoothing method proposed by Breithaupt *et al.* [20]. The superiority of CEM over these [7] is the reason why—for the sake of conciseness—we mostly concentrate on **CEM** as baseline in this work except for the final results, where we also present the results of a traditional speech enhancement system using the **DD** approach as *a priori* SNR estimator. As a more recent approach we also test against a DNN-based ideal ratio mask (**IRM**) solution. Furthermore, our recently proposed method [8], dealing with the enhancement of the spectral envelope, dubbed cepstral envelope estimation (**CEE**), is now also used as a baseline. It is the counterpart of the CEM approach and has shown to further improve CEM, when combined in a serial manner, where first CEE is applied followed by CEM. For more details we kindly refer to [8], where we also show that the baselines are able to compete with modern end-to-end speech enhancement techniques such as the ideal ratio mask [2], [5]. This serial combination is also used as further baseline, named **CEE → CEM**.

Both solo approaches, CEE and CEM, are depicted in Fig. 1, where switches S_1 and S_2, both in upper position, represent the CEM approach, and both in lower position, represent the CEE approach[1]. As can be seen in Fig. 1, both methods share a common pipeline up to the LPC analysis, where it branches to facilitate the enhancement of each component, excitation and envelope, separately. The use of the source-filter model allows to split the enhancement task into two sub-problems which are briefly revised as follows.

A. Cepstral Excitation Manipulation (CEM)

The baseline configuration of the **CEM** approach [7] is depicted in more detail in Fig. 2 with switch S_3 in upper position. The first block (Feature Conversion) represents a feature transformation from the spectral domain to the cepstral domain by applying a discrete cosine transform of type II (DCT-II), followed by a simple pitch estimation algorithm [32]. The quefrency bin index m_{F_0} corresponding to the pitch frequency is estimated by selecting the quefrency bin in a certain range of fundamental frequency-representing bins, that exposes the highest amplitude. Following, a pretrained clean speech excitation template $\hat{c}_\ell^R(m)$ that depends on the estimated fundamental frequency is selected from a storage and used further. The following processing aims to adjust the energy of the synthesized excitation signal by replacing the amplitude of the template's zeroth coefficient $\hat{c}_\ell^R(0)$ by the amplitude representing the energy of the preliminary enhanced residual signal $c_\ell^R(m)$ by

$$\hat{c}_\ell^R(0) = c_\ell^R(0). \tag{6}$$

[1] As a further option it is possible to apply CEM and CEE in parallel, when switch S_1 is in upper, and switch S_2 is in lower position. This parallel approach has been evaluated in [8] and shown to improve the noise attenuation. However, it also affects the speech component quality compared to the solo approaches CEM or CEE, and thus is disregarded here.

https://doi.org/10.24355/dbbs.084-202102121510-0

A further step to enhance the excitation signal is that the incoming amplitude of the quefrency bin that represents the fundamental frequency $c_\ell^R(m_{F_0})$ is overestimated by a factor $\alpha > 1$ in order to boost the harmonic structure and simultaneously lower the energy between the harmonics to obtain a higher noise attenuation. It is then also inserted into the template as

$$c_\ell^{\hat{R}}(m_{F_0}) = \alpha \cdot c_\ell^R(m_{F_0}). \tag{7}$$

After these manipulation steps, the cepstral vector is transformed back into the spectral domain by an inverse DCT-II, yielding the manipulated residual spectral amplitude $|\hat{R}_\ell(k)|$. By using a cepstral representation of the excitation signal, one is able to address and manipulate all harmonics in the signal's spectral representation at a single cepstral bin.

Employing the F_0 estimate, finally some start/end decay to the spectral representation is applied, as this ensures a somewhat more natural rise and decay of the harmonic structure which might have been corrupted by the manipulations or is erroneous in the templates itself. The start decay is a simple linear continuation of the rising edge for the first harmonic while the end decay is applied in the same manner to the last fully representable harmonic, but in this case to the declining edge. Both measures lead to an attenuation of spectral content prior to the first and after the last harmonic, where no speech content is expected (further details in [7]).

B. Cepstral Envelope Estimation (CEE)

The counterpart of CEM is the enhancement of the spectral envelope which has been extensively investigated in [8], dubbed cepstral envelope estimation (**CEE**). We will briefly introduce the optimal solution in the following. The general idea (see also [33]–[35]) is to find a mapping between the spectral envelope of the preliminary denoised signal and a linear combination of pretrained N-dimensional prototypes $\tilde{\boldsymbol{c}}_i^H = [\tilde{c}_i^H(1), \ldots, \tilde{c}_i^H(m), \ldots, \tilde{c}_i^H(N)]^{\mathrm{T}}$, obtained from clean speech recordings which are stored in a codebook $\mathcal{C} = \{\tilde{\boldsymbol{c}}_i^H\}$. The prototypes are indexed by $i \in \mathcal{N}_S = \{0, 1, 2, \ldots, N_S\}$, where $i = 0$ represents a prototype for non-speech frames. The advantages of a cepstral representation are used once more, with the difference that not the DCT-II is used, but the LPC coefficients $a_\ell(m)$ are transformed directly by the recursive formula from [36], to obtain the cepstral representation $c_\ell^H(m)$. This allows to work safely with the coefficients without risking any instabilities of the filter as would be the case when working on LPC coefficients directly. A codebook size of $N_S + 1 = 65$ has proven to be optimal with dimensionality $N = 10$ and a simple feedforward classification DNN consisting of six hidden layers and 58 nodes each. It was shown, that the sigmoid activation functions have lead to slightly higher accuracies than rectified linear units and a softmax output layer. The network's input is the cepstral representation $c_\ell^H(m)$ and the output can be understood as a probability distribution over the prototypes in the codebook as

$$\mathrm{P}(s_\ell = i | \boldsymbol{x} = \boldsymbol{o}_\ell). \tag{8}$$

Hereby, s_ℓ represents a hidden state which is a proxy for the unknown truth behind the observation, i.e., the true clean spectral envelope, while the corresponding observation is defined as $\boldsymbol{o}_\ell = \left[c_\ell^H(1), \ldots, c_\ell^H(N)\right]$. Having obtained the probability distribution, MMSE estimation is performed by

$$c_\ell^{\hat{H}}(m) = \sum_{i \in \mathcal{N}_S} \mathrm{P}(s_\ell = i | \boldsymbol{x} = \boldsymbol{o}_\ell) \cdot \tilde{c}_i^H(m), \tag{9}$$

and the estimated cepstral vector $\boldsymbol{c}_\ell^{\hat{H}}$ is converted back to the estimated envelope spectral amplitudes $|\hat{H}_\ell(k)|$ by applying an IDCT-II. Further details can be found in [8].

C. Decision-Directed Approach (DD)

Originally proposed by Ephraim and Malah in [31], the decision-directed (**DD**) approach is still considered as an important baseline. Even though the previously mentioned baselines already outperform the DD approach, many researchers are also interested to see improvement vs. a speech enhancement system using the DD *a priori* SNR estimator. We use the DD estimator with $\beta_{\mathrm{DD}} = 0.975$ and $\xi_{\min} = -15$ dB to prevent too many musical tones.

D. Ideal Ratio Mask (IRM)

As a more recent approach we also test against an IRM approach based on a feedforward DNN which is in line with [2], [5]. The network consists of three hidden layers with 1024 nodes each and rectified linear units as activation functions. The total amount of parameters is 2,364,545. We are using log-spectral amplitude input features and calculate the target gains for training as

$$G_\ell^{\mathrm{IRM}}(k) = \left(\frac{|S_\ell(k)|^2}{|S_\ell(k)|^2 + |D_\ell(k)|^2} \right)^\beta, \tag{10}$$

with $\beta = 1.0$. In fact, this spectral weighting rule ($\beta = 1.0$) has been used for learning a lookup table with spectral gains based on the *a priori* and *a posteriori* SNR before [29].

IV. DNN-Supported Cepstral Excitation Manipulation

Incorporating the novel opportunities of deep learning we want to explore the potential of the CEM idea when it is realized by a neural network instead of the classical signal processing measures that have been applied until now (see Section III-A). We show both approaches in Fig. 2, where the classical baseline CEM is depicted in the upper path (switch S_3 in upper position) and the new proposed approach, dubbed **CEM-DNN**, in the lower path (switch S_3 in lower position). As further option a smooth start and end decay can be applied to the manipulated amplitude spectrum of the residual signal (S_4 in upper position), to ensure smooth transitions which was necessary for the template-based CEM approach. The start and end decay function still relies on the simple F_0 estimator proposed by [32], however, this is a less critical application compared to the former selection of templates based on the same estimate in state-of-the-art CEM. Following Fig. 2, the feature conversion block (see also Fig. 3) transforms the log-spectral amplitudes of the residual signal

https://doi.org/10.24355/dbbs.084-202102121510-0

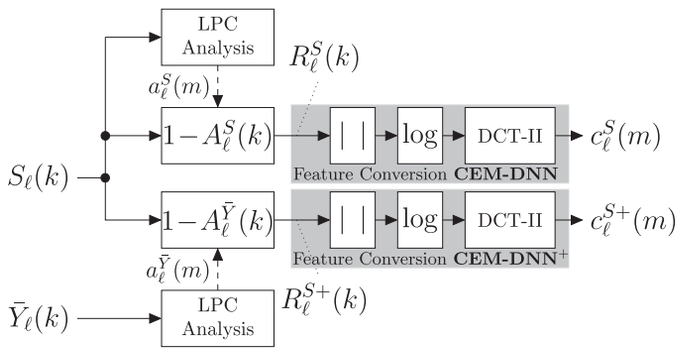


Fig. 3. Block diagram of the processing pipeline for two different representations of training targets for the **CEM-DNN** and **CEM-DNN**$^+$ approaches.

$R_\ell(k)$ into the cepstrum by applying the DCT-II, resulting in $c_\ell^R(m)$. When we apply normalization, all data is processed by bin-wise cepstral mean and variance normalization in order to remove potential channel mismatches. *Note that the core difference to the classical CEM approach is the replacement of the excitation templates $\tilde{\mathbf{c}}_i^H$ and MMSE estimation (9), as well as of the two manipulations (6) and (7) by a regression DNN.* In consequence, the core of CEM-DNN also does not need an F_0 estimator any more.

The output $c_\ell^{\hat{R}}(m)$ of the DNN is rescaled if necessary and subsequently transformed back into the spectral domain by the IDCT-II and optionally the start/end decay is applied. We finally obtain the estimated spectral amplitudes of the residual signal $|\hat{R}_\ell(k)|$. Rescaling of the DNN output is performed by using the mean and variance obtained from the respective data set. This translates to a practical system, as noise reduction is an uplink feature, which allows to calculate the required mean and variance of the signals after the preliminary noise reduction and LPC analysis for the input of the DNN, or during good SNR conditions to rescale the output of the DNN. In the following we will introduce our general setup for the DNN training and two different kinds of target representations.

A. DNN Training

The general setup of our DNN training process is based on the `KERAS` toolkit [37] together with the `TensorFlow` [38] backend. We normalize all *input features* by cepstral mean and variance normalization and in some cases we also normalize the target representation. The normalization is important to provide similar data ranges to the network which can ensure convergence and stability during training [39]. A similar argument holds for *target* normalization when regression networks are used: We explore the benefits of target normalization in more detail in Section VI-B2, however, it is not always applicable. Each input layer has the same amount of nodes as the input feature dimension $N = 256$. The subsequent N_H hidden layers each have N_N nodes. As we have experienced before, the difference between sigmoid and rectified linear units as activation function can be very marginal [8]. Since we did not encounter any problems with vanishing gradients so far, but obtained slightly better results with sigmoid activation functions, we decided to only investigate sigmoid activation functions in this case. The final output layer has also $N = 256$ nodes and uses linear activation functions since we only investigate regression DNNs. The parameters of the network, the biases and weights, are all initialized as proposed by Glorot *et al.* [40]. We employ the mean squared error (MSE) loss function in order to make the network learn the mapping between input and output representations. The training data is randomly accessed by the sequencing mechanism and provides batches of $L = 2048$ input and target frames at a time. For the gradient-based optimization we use the adaptive moment estimator (Adam) [41] with default parameterization, including a learning rate of $\eta = 0.001$. The networks are trained straight for 300 epochs from which the best model on some development set is selected and used further. In the following, we describe the two types of target representations and their advantages and disadvantages.

B. Target Representations

Since we aim to improve the excitation signal, the intuitive way is to simply extract excitation signals $R_\ell^S(k)$ from clean speech data $S_\ell(k)$ as targets for the training process of the DNN. The corresponding input features are the noisy, or in our case the already preliminary denoised, residual signals obtained from multiple simulated SNR and noise conditions. The pipeline for the target extraction is shown in Fig. 3 at the top. The frequency-domain representation of the clean speech data $S_\ell(k)$ is used for LPC analysis and subsequently filtered with the corresponding analysis filter $1 - A_\ell^S(k)$. The resulting spectral representation of the residual signal $R_\ell^S(k)$ is then subject to feature conversion, i.e., conversion to the log-amplitude spectrum, followed by the DCT-II to obtain the cepstral coefficients $c_\ell^S(m)$. The advantage of this target representation is that it is possible to obtain mean and variance data of $c_\ell^S(m)$ for the rescaling of the DNN output during inference (it is sufficient to collect these statistics from time to time during good SNR conditions), even in a practical application. Note that in such a practical implementation the input $S_\ell(k)$ for *both* the LPC analysis and the LPC analysis filtering in the upper path of Fig. 3 would have to be replaced by $\bar{Y}_\ell(k)$. In Fig. 1 it can be seen that the estimated amplitudes of the residual signal $|\hat{R}_\ell(k)|$ are mixed with the envelope of the preliminary denoised signal $|H_\ell(k)|$ (switches S_1, S_2 in upper positions). Thus, there will be still some mismatch between residual and envelope. We refer to the CEM method trained with these particular targets in the following as **CEM-DNN**.

Better targets for the training can be obtained by also considering the preliminary denoised signal's envelope. This is shown in Fig. 3 at the bottom, where the LPC coefficients are obtained from the preliminary denoised signal $\bar{Y}_\ell(k)$. The clean speech signal $S_\ell(k)$ is then filtered with the corresponding analysis filter $1 - A_\ell^{\bar{Y}}(k)$ which yields, after the usual feature conversion, the cepstral coefficients $c_\ell^{S+}(m)$ of our other target features. Those features allow, theoretically, the reconstruction of the clean speech signal even with a preliminary denoised signal's envelope during inference. However, the required mean and variance data of $c_\ell^{S+}(m)$ for the rescaling of the network's output can only be obtained in lab conditions, since the core idea of this approach consists of the discrimination between $S_\ell(k)$ and $\bar{Y}_\ell(k)$, and the use of both. This prohibits target

https://doi.org/10.24355/dbbs.084-202102121510-0

normalization in practice, or target normalization is done on some static precalculated mean and variance from, e.g., the training data. The corresponding CEM approach, using mean and variance obtained in lab conditions, is dubbed **CEM-DNN$^+$**.

V. Experimental Setup

In the following, we describe the used databases for the development process of our system and also the instrumental quality measures which are used for the final evaluation of the baselines and the proposed approach.

A. Databases

In order to ease comparison to our earlier works [8], we use the same database setup for training, development, and testing. The training and development sets are based on the TIMIT database [42], where the training set is used as training set and the test set of the TIMIT database as development set for our experiments. We finally report results on the NTT super wideband database [43] (only British and American English speakers) which serves as a test set and allows us to also report cross-database results. The clean databases are corrupted by noises from the QUT [44] and the ETSI [45] databases. Please note that all data is downsampled to 8 kHz for our experiments. Except for the male single voice distractor noise file from the ETSI database, all files are used. Among them we find, e.g., babble, road, car, office, aircraft, and also kitchen noise. Four noise files are reserved for a special test set with unseen noise files, which is important to show how well results of data-driven algorithms generalize to unseen data. We generate noisy data at 8 kHz sample rate for six SNR conditions, i.e., –5 dB to 20 dB in steps of 5 dB. The noise files are split up into non-overlapping sections, where 60% are used for training, 20% for development, and the remaining 20% for testing. Each file from the two speech databases is mixed with a random part of each noise file's respective section (four noise files held out for test with unseen noise files, as said above). To accomplish this, both clean speech signal and noise part, are level-adjusted according to ITU-T P.56 [46] and subsequently superimposed. In total we generate 6 (SNRs) $\times$ 53 (noise files) $=$ 318 conditions, represented by 318 $\times$ 4620 $=$ 1,469,160 (training set) and 318 $\times$ 1680 $=$ 534,240 (development set) noisy speech files[2]. Last, our framing setup is using a periodic square root Hann window, along with a frame shift of 50% and a frame length of $K = 256$ samples.

[2]This is a multitude of files that forces us to develop strategies to successfully cope with a huge amount of data for the training, and also the development process. Due to the large amount of data, i.e., the input features $c_\ell^R(m)$ and the targets $c_\ell^{S+}(m)$ for all 318 conditions, consuming together around 532 GB of disk space when stored as single-precision floating-point values, we decided to take two measures: First, we store all data as half-precision floating-point values resulting in a reduction to 266 GB and second, for our development process of the network structure, we optimize on the -5 dB SNR condition for all noise types only which reduces the amount of data further to roughly 44 GB. This allows us to be more flexible and we finally show that the loss-optimized topology found by single SNR condition training is also optimal for the multi-condition training which takes much more time.

B. Instrumental Quality Measures

As basis for our evaluation we employ the white-box approach [47], which allows us to assess the speech and noise component quality separately (see also ITU-T P.1100 Section 8 [48]). This is achieved by applying the gain function $G_\ell(k)$ not only to the microphone signal $Y_\ell(k)$, in order to obtain the enhanced signal $\hat{S}_\ell(k)$, but also to the separate components. This yields the *filtered* clean speech component $\tilde{S}_\ell(k) = S_\ell(k) \cdot G_\ell(k)$, and the *filtered* noise component $\tilde{D}_\ell(k) = D_\ell(k) \cdot G_\ell(k)$.

Both are subsequently transformed into the time domain by applying an inverse DFT followed by overlap-add synthesis, resulting in $\tilde{s}(n)$ and $\tilde{d}(n)$, respectively.

For the instrumental evaluation of the approaches, we use measures of two different categories in order to assess the amount of noise attenuation on the one hand and speech quality and intelligibility on the other hand. For the first, we use the segmental noise attenuation (NA_{seg}) measure [29] which can be obtained as

$$\text{NA}_{\text{seg}} = 10 \log_{10} \left[\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{NA}(\ell) \right], \tag{11}$$

with

$$\text{NA}(\ell) = \frac{\sum_{\nu=0}^{N-1} d(\nu + \ell N)^2}{\sum_{\nu=0}^{N-1} \tilde{d}(\nu + \ell N + \Delta)^2}.$$

The measure depicts the logarithmic average over the noise attenuation of all frames $\ell \in \mathcal{L}$. Each frame contains $N = 256$ samples and Δ compensates potential processing delay. A high value indicates good performance.

As additional measure to assess the SNR improvement on a global level we introduce the delta SNR which is calculated as

$$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}. \tag{12}$$

SNR_{out} represents the SNR of the *filtered* speech and noise component after processing and SNR_{in} the corresponding SNR of the clean speech and noise signals.

The speech quality of the *filtered* speech component $\tilde{s}(n)$ is measured by the PESQ score (mean opinion score, listening quality objective (MOS-LQO)) [26], [27] with $s(n)$ as the reference signal.

As fourth measure, we use the short-time objective intelligibility measure (STOI) [49] to rate the intelligibility of the enhanced speech signal $\hat{s}(n)$ compared to the clean speech signal $s(n)$. The closer the value is to unity, the better.

VI. Evaluation and Discussion

A. Oracle Experiments and Motivation

First of all, we conduct two oracle experiments which serve as motivation for our research. In Figs. 4 and 5, both oracle experiments show the performance of an *a priori* SNR estimator with different use of partial oracle knowledge, set in the context of the noise reduction framework as described in Section II-B. They use the same noise power estimate obtained by MS, along with an adjusted numerator as follows: The *oracle excitation* experiment (solid purple line, diamond markers) mixes the denoised

https://doi.org/10.24355/dbbs.084-202102121510-0

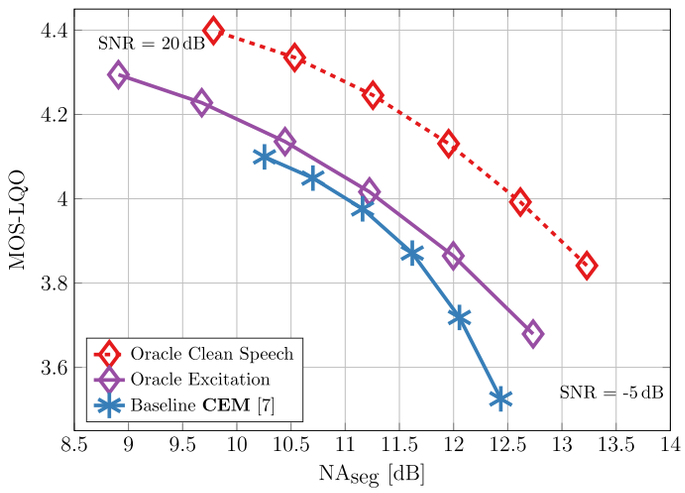


Fig. 4. Two oracle experiments showing the motivation and the unexhausted potential of the baseline **CEM** approach in terms of NA_{seg} and speech component quality measured by speech component **MOS-LQO**. All results are obtained on the **development set**.

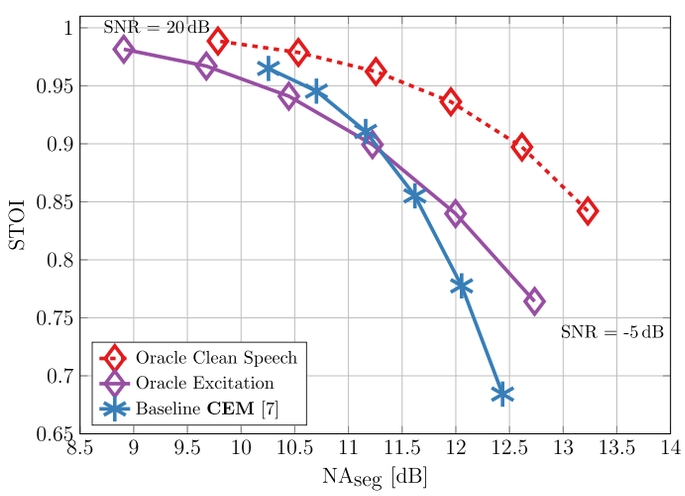


Fig. 5. Two oracle experiments showing the motivation and the unexhausted potential of the baseline **CEM** approach in terms of NA_{seg} and speech intelligibility measured by **STOI**. All results are obtained on the **development set**.

envelope $|H_\ell(k)|$ (see Fig. 1) with the oracle excitation signal obtained from clean speech. A more advanced oracle experiment (dashed red line, diamond markers) uses the *oracle clean speech* in the numerator for the *a priori* SNR, which assumes to know not only the clean speech excitation but also the corresponding clean speech envelope. The results show quite expected behavior, as with increasing oracle knowledge the potential gain in NA_{seg}, MOS-LQO, and also STOI increases, compared to the baseline **CEM** approach (solid blue line, asterisk markers). In the figures, each marker depicts a certain SNR condition from -5 dB at the bottom, in steps of 5 dB, up to 20 dB at the top. Using the oracle excitation signal shows less potential in terms of NA_{seg} compared to using the oracle clean speech signal. However, the potential gain in speech component quality and intelligibility (the vertical in both figures) is still worth pursuing, especially when considering the low-SNR conditions.

TABLE I
EVALUATION OF THE **MSE** LOSS FOR VARIOUS NETWORK TOPOLOGIES BASED ON THE **-5 DB SNR** CONDITION WITH $c_\ell^S(m)$ TARGETS FOR THE **DEVELOPMENT SET**

N_H	$N_N = 64$	$N_N = 128$	$N_N = 256$	$N_N = 512$	$N_N = 1024$
1	0.839	0.811	0.796	0.787	0.782
2	0.830	0.794	0.771	0.758	0.753
3	0.823	0.785	0.761	0.746	0.742
4	0.816	0.779	0.754	0.742	0.742
5	0.815	0.776	0.752	0.741	0.742
6	0.813	0.774	0.750	**0.739**	0.744

TABLE II
EVALUATION OF THE **MSE** LOSS FOR VARIOUS NETWORK TOPOLOGIES BASED ON **ALL SNR** CONDITIONS WITH $c_\ell^S(m)$ TARGETS FOR THE **DEVELOPMENT SET**

N_H	$N_N = 64$	$N_N = 128$	$N_N = 256$	$N_N = 512$	$N_N = 1024$
1	0.744	0.679	0.643	0.654	0.648
2	0.732	0.661	0.632	0.622	0.615
3	0.726	0.654	0.631	0.608	0.604
4	0.721	0.648	0.632	0.603	**0.600**
5	0.719	0.645	0.633	0.601	0.601
6	0.718	0.643	0.636	**0.600**	0.603

B. Cepstral Excitation Manipulation With DNN

In order to tap the potential of the cepstral excitation manipulation approach we decide to integrate a regression DNN. We briefly scanned on the development set through various parameters and ended up with the configuration as given in Section IV-A, as results stayed quite comparable. However, the topology of the network had quite some impact on the quality of the network's output. In Table I we show the MSE loss for several configurations of hidden layers N_H and their number of nodes N_N for the -5 dB SNR condition of the development set. It was necessary to make optimizations on a small set of data as the training process with all SNR conditions is quite time-consuming. In Table II (all SNR conditions), the MSE loss appears to be comparable for $N_N \in \{512, 1024\}$, which is natural since due to mean and variance normalization of the targets the number range of the loss also decreases. Solving the tie in Table II, we feel comfortable to put focus on the -5 dB condition (Table I) and decide for a configuration of $N_H = 6$ and $N_N = 512$ resulting in a total amount of 1,576,192 parameters. It might be possible that with increasing number of hidden layers the loss would drop further, which we expect to be rather marginal in this case. Note that the trainings have been conducted with $c_\ell^S(m)$ targets and we assume that the results translate also to $c_\ell^{S+}(m)$ targets without significant aberrations.

Now, we investigate the influence of the applied start and end decay as depicted in Fig. 2, the effects of target normalization, and the two different types of target representations as shown in Fig. 3.

1) Influence of Start and End Decay Function: In Figs. 6 and 7, we depict the **CEM-DNN** approach (square markers) which aims at estimating the clean excitation signal and the **CEM-DNN$^+$** approach (plus markers) which aims at compensating also for the denoised spectral envelope, and thus to obtain the

https://doi.org/10.24355/dbbs.084-202102121510-0

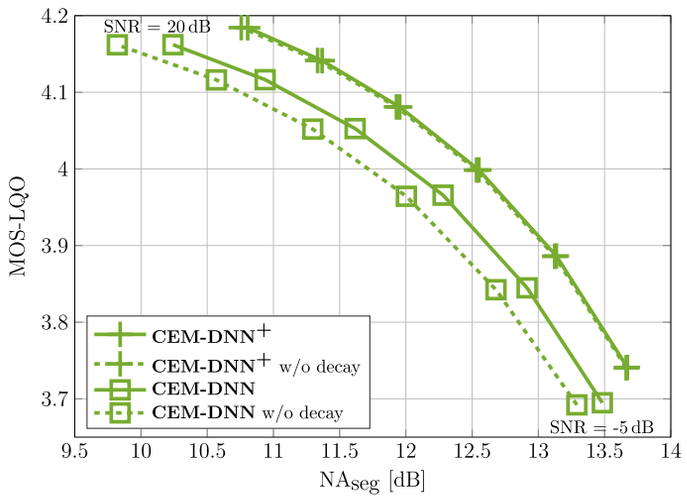


Fig. 6. The effect of applying start and end decay to either **CEM-DNN** or **CEM-DNN**$^+$ measured by NA$_{seg}$ and speech component **MOS-LQO** on the **development set**.

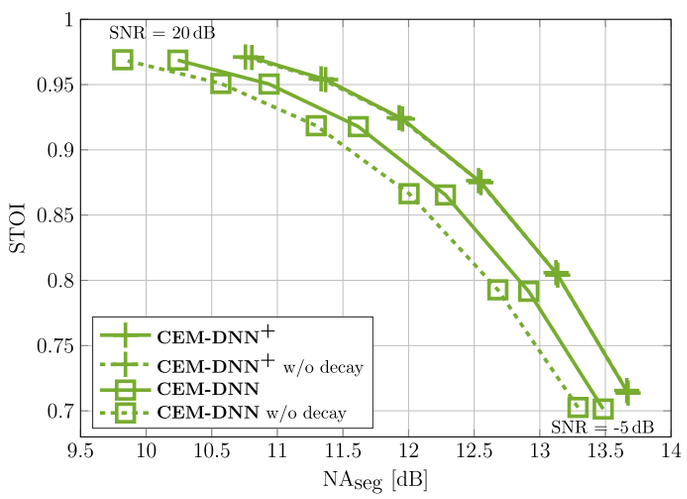


Fig. 7. The effect of applying start and end decay to either **CEM-DNN** or **CEM-DNN**$^+$ measured by NA$_{seg}$ and **STOI** on the **development set**.

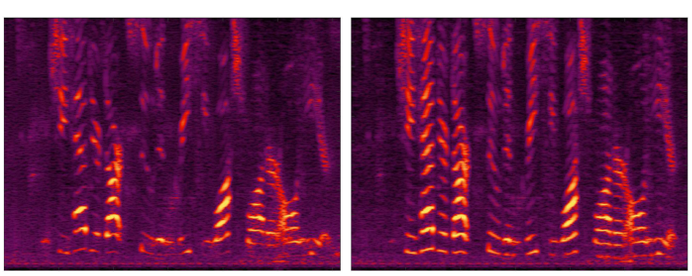


Fig. 8. **Spectrograms** of an enhanced microphone signal from the **development set** at 10 dB SNR with `CAFE-CAFE-1` noise processed by **CEM-DNN** trained **without** (left) and **with mean/variance normalization** (right) of the targets.

clean speech signal. Both approaches are depicted with applied start and end decay (solid green lines) and without (dashed green lines). The results show that the start and end decay has only an effect on **CEM-DNN** while the effect on **CEM-DNN**$^+$ is

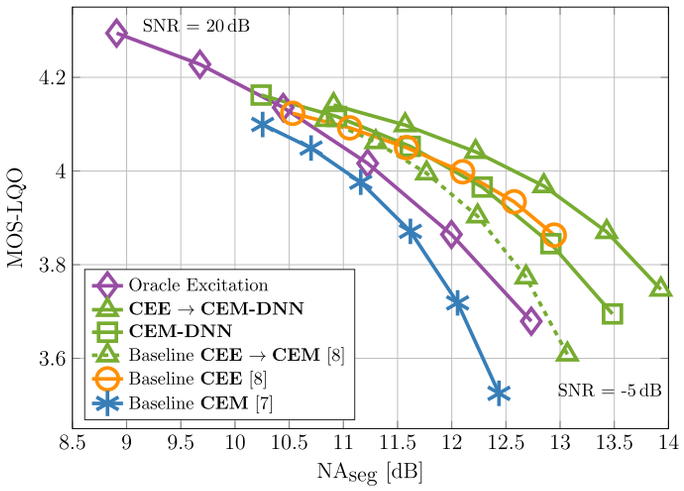


Fig. 9. Showing the performance (NA$_{seg}$ and speech component **MOS-LQO**) on the **development set** for the baseline approaches, the new **CEM-DNN** method with applied decay, its serial concatenation with **CEE**, and the oracle experiment depicting the upper limit of the CEM approach.

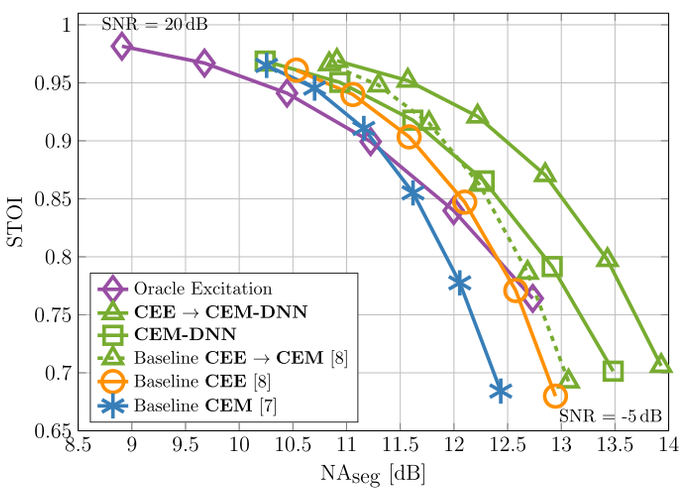


Fig. 10. Showing the performance (NA$_{seg}$ and **STOI**) on the **development set** for the baseline approaches, the new **CEM-DNN** method with applied decay, its serial concatenation with **CEE**, and the oracle experiment depicting the upper limit of the CEM approach.

negligible. This is quite interesting, as it indicates that the application of the start and end decay might be naturally attributed to the envelope and is automatically compensated for by the DNN. Furthermore, the results show that **CEM-DNN** is able to benefit from the application of the start and end decay as NA$_{seg}$ is consistently improved without significant impact on MOS-LQO and STOI. From here on all experiments are shown with applied start and end decay function.

2) Influence of Target Normalization: Next, we investigate the effect of target normalization in Fig. 8, showing the spectrograms of an enhanced microphone signal from the development set with `CAFE-CAFE-1` noise and 10 dB SNR condition. The microphone signal is then processed by **CEM-DNN** with applied start and end decay, once for a network trained without (left spectrogram), and once for a network trained with (right spectrogram) target normalization. The richness of the spectrogram

https://doi.org/10.24355/dbbs.084-202102121510-0

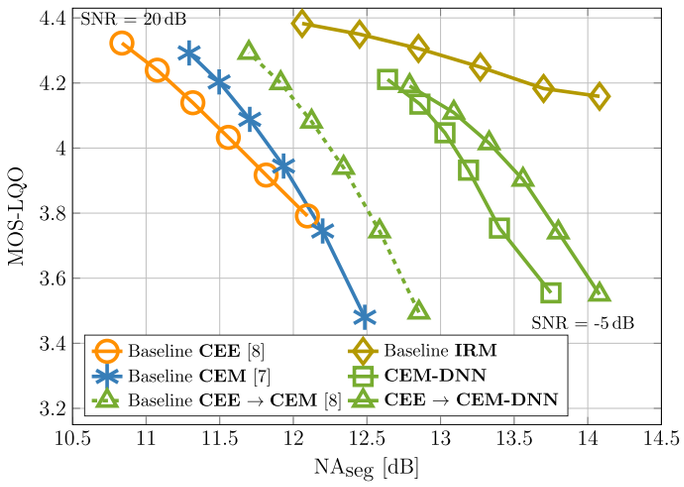


Fig. 11. Showing the performance (NA$_{seg}$ and speech component **MOS-LQO**) on the **test set** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

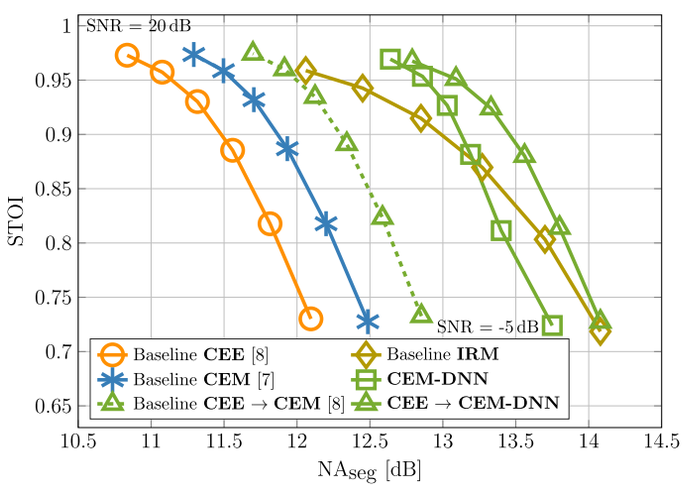


Fig. 12. Showing the performance (NA$_{seg}$ and **STOI**) on the **test set** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

on the right shows the importance of target normalization which results in a much better preservation, especially in the lower frequency regions, of harmonic structures compared to the left spectrogram. This is a problem for the **CEM-DNN**$^+$ approach, as rescaling of the DNN output, as mentioned in Section IV-B, during inference would only be possible with pre-trained statistics, without any possibility of adaptation. Hence, we will continue only with **CEM-DNN**, with start and end decay, and with target normalization.

3) Results for the Development Set: In Figs. 9 and 10 we show the performance of the baselines **CEM** (solid blue line, asterisk markers), **CEE** (solid orange line, circle markers), and the serial concatenation of the former two approaches **CEE** → **CEM** (dashed green line, triangle markers) on the development set. Furthermore, we show the upper limit of the **CEM** approach by using the oracle excitation (solid purple line, diamond markers), the new approach **CEM-DNN** with start and end decay,

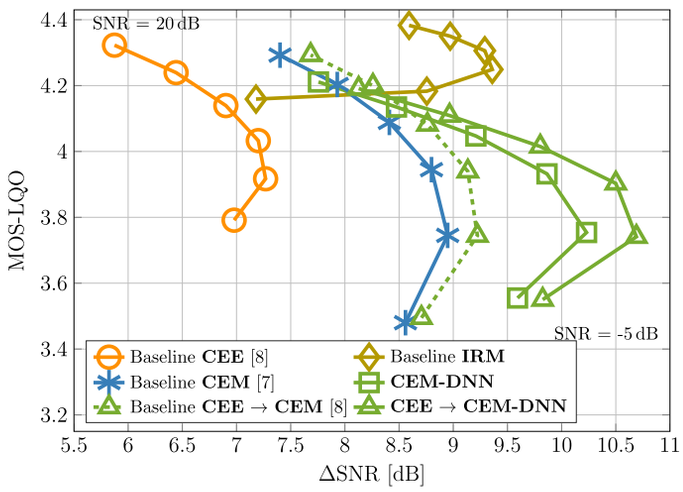


Fig. 13. Showing the performance (ΔSNR and speech component **MOS-LQO**) on the **test set** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

and its serial concatenation with the baseline **CEE**, labelled as **CEE** → **CEM-DNN** (solid green line, triangle markers). The noise attenuation of **CEM-DNN** improves over **CEM** by up to 1 dB for the −5 dB SNR condition, while increasing MOS-LQO by more than 0.1 points and also slightly improving STOI. This is an absolute improvement for the worst and most important SNR condition. The approach is even able to outperform **CEE** → **CEM** consistently up to and including the 5 dB SNR condition. The **CEE** approach still shows superior speech component quality measured by MOS-LQO, however, it is unable to remove noise between the harmonics and falls behind in most conditions for NA$_{seg}$ and also slightly for STOI. Surprisingly, compared to the oracle excitation experiment, **CEM-DNN** obtains higher NA$_{seg}$, and in some cases comparable MOS-LQO, but does not match in speech intelligibility. In serial combination with the **CEE** approach, **CEE** → **CEM-DNN** yields further absolute improvement in terms of NA$_{seg}$ by up to more than 0.5 dB with comparable MOS-LQO and STOI values.

4) Results for the Test Set: On the test set, which evaluates a different database, shown in Figs. 11–13, the behavior is quite similar. **CEM-DNN** and also **CEE** → **CEM-DNN** obtain higher NA$_{seg}$ by more than 1 dB over their corresponding baseline. Thereby, MOS-LQO is slightly improving for the −5 dB SNR condition and STOI stays about the same. Only in high-SNR conditions the proposed approaches drop slightly in speech component quality, which is, however, uncritical as the quality still remains very high and STOI also reports no significant loss of intelligibility.

In addition to that, we also show the **IRM** baseline (solid sand line, diamond markers) which shows exceedingly high speech component quality. However, in terms of NA$_{seg}$ and STOI the approach falls behind **CEE** → **CEM-DNN** with increasing SNR. In Fig. 13, for low and medium SNRs, the SNR improvement (ΔSNR) of the IRM approach falls far behind the proposed approach which outperforms all other approaches consistently. This also indicates that the attenuation characteristic of IRM is

https://doi.org/10.24355/dbbs.084-202102121510-0

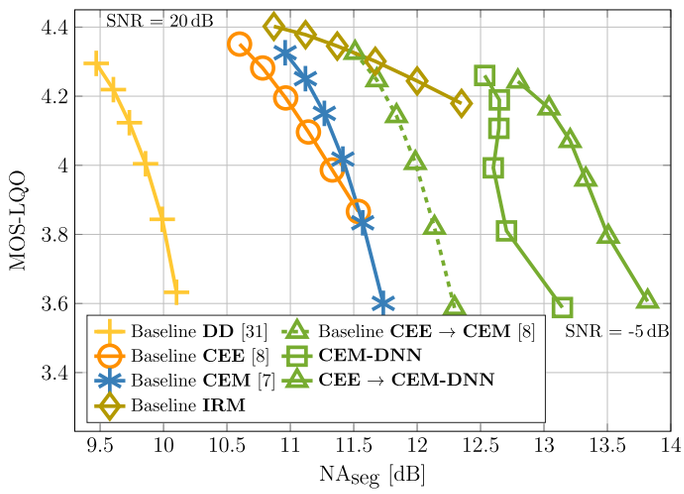


Fig. 14. Showing the performance (NA$_{seg}$ and speech component **MOS-LQO**) on the **test set with unseen noise files** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

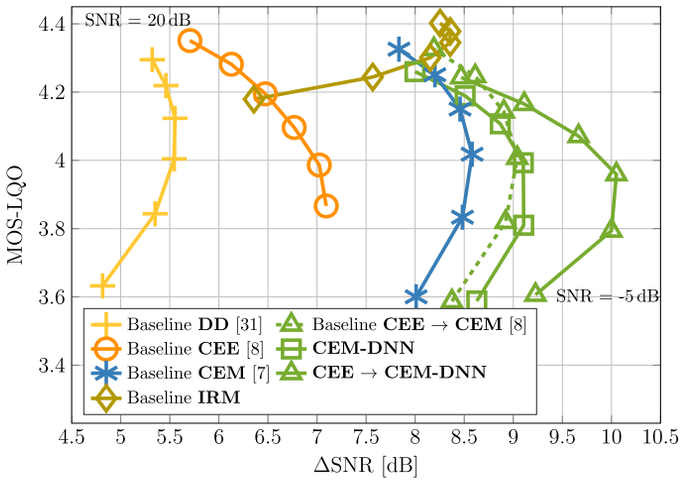


Fig. 16. Showing the performance (ΔSNR and speech component **MOS-LQO**) on the **test set with unseen noise files** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

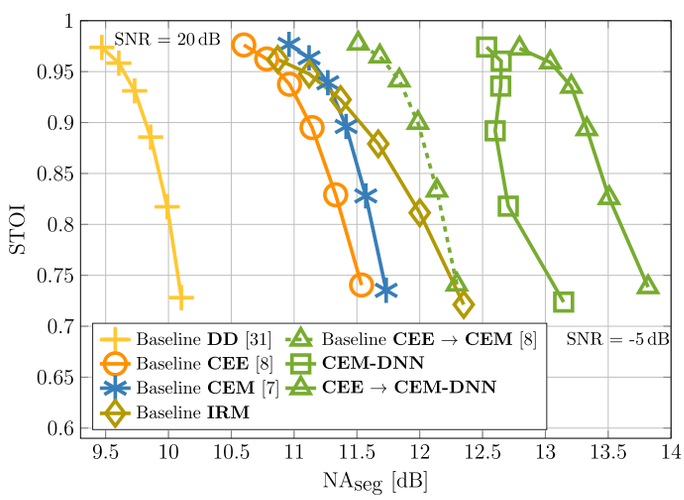


Fig. 15. Showing the performance (NA$_{seg}$ and **STOI**) on the **test set with unseen noise files** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

more broadband and thus affecting speech and noise simultaneously, which explains the high MOS-LQO as PESQ is internally adjusting the level. Another issue with the **IRM** approach is the mentioned discontinuity problem as detailed in [3], and also observed in [8, Section VI-B2].

5) Results for the Test Set With Unseen Noise Files: The final evaluation on the test set with completely unseen noise files[3] during training is shown in Figs. 14–16. The results show that the performance transfers quite nicely to (also non-stationary) noise files that have not been seen during training, which is closest to a real-world scenario. Except for the already explained high speech component MOS-LQO, the proposed approach outperforms IRM clearly. Analyzing MOS-LQO and STOI at −5 dB SNR in Figs. 14 and 15, we observe an 1.5 dB NA$_{seg}$ advantage of **CEE → CEM-DNN** vs. **IRM**, which is not the case in Figs. 11 and 12 (seen noises). This shows that baseline **IRM** generalizes not as good w.r.t. background noises. Compared to the respective baselines, there is no significant trade-off for speech intelligibility, and for the speech component quality only minor drawbacks in the high-SNR conditions, where the absolute speech component quality is already very high (above 4 MOS-LQO points).

For similar MOS-LQO (Fig. 14) and STOI (Fig. 15) we can also report a gain in NA$_{seg}$ of approximately 1.5 dB for the −5 dB SNR condition (lowest marker) when comparing the new **CEM-DNN** to the baseline **CEM**, and also when comparing **CEE → CEM-DNN** to the baseline **CEE → CEM**. Compared to the **DD** approach, the proposed **CEE → CEM-DNN** obtains more than 3 dB NA$_{seg}$ while maintaining comparable speech component quality and speech intelligibility[4].

VII. Conclusion

In this work we have investigated the application of a deep neural network (DNN) to the cepstral excitation manipulation (CEM) approach for *a priori* SNR estimation in a speech enhancement task. We have investigated two target representations, where one is not applicable to practical systems and the other shows convincing performance. Furthermore, we could verify the benefit of applying some start and end decay to the estimated residual signal and have shown the importance of target normalization. Thus, we have successfully enhanced the classical signal processing-based CEM approach by introducing a simple feedforward DNN which has lead to an *improvement on unseen and non-stationary noise files by up to 1.5 dB of segmental noise attenuation without sacrificing speech component quality and speech intelligibility*. Compared to a traditional speech

[3] `Fullsize_Car1_80Kmh`, `Outside_Traffic_Crossroads`, `Pub_Noise_Binaural_V2`, `Work_Noise_Office_Callcenter`

[4] Audio samples can be found under: https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2019-taslp-cem/

https://doi.org/10.24355/dbbs.084-202102121510-0

enhancement system with the decision-directed *a priori* SNR approach, an *improvement of even more than 3 dB segmental noise attenuation with comparable speech intelligibility is achieved on the same data.*

REFERENCES

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[2] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, Apr. 2015, pp. 4390–4394.

[3] S.-W. Fu, T.-W. Wang, Y. Taso, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[4] S. Mirsamadi and I. Tashev, "Causal speech enhancement combining data-driven learning and suppression rule estimation," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 2870–2874.

[5] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[6] F. Bao and W. H. Abdulla, "A new ratio mask representation for CASA-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 7–19, Jan. 2019.

[7] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous a priori SNR estimation by cepstral excitation manipulation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[8] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "DNN-supported speech enhancement with cepstral estimation of both excitation and envelope," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2460–2474, Dec. 2018.

[9] S. Srinivasan and J. Samuelsson, "Speech enhancement using a-priori information," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1405–1408.

[10] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[11] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[12] T. Rosenkranz, "Noise codebook adaptation for codebook-based noise reduction," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Tel Aviv, Israel, Aug. 2010.

[13] Q. He, F. Bao, and C. Bao, "Multiplicative update of auto-regressive gains for codebook-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 457–468, Mar. 2017.

[14] Y. Yang and C. Bao, "DNN-based AR-wiener filtering for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 2901–2905.

[15] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[16] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[18] I. Cohen, "Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation," *Speech Commun.*, vol. 47, no. 3, pp. 336–350, Nov. 2005.

[19] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[20] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Mar. 2008, pp. 4897–4900.

[21] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori SNR estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[22] S. Elshamy, N. Madhu, W. J. Tirry, and T. Fingscheidt, "An iterative speech model-based a priori SNR estimator," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 1740–1744.

[23] L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, "Convex combination framework for a priori SNR estimation in speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 4975–4979.

[24] J. Stahl and P. Mowlaee, "A simple and effective framework for a priori SNR estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5644–5648.

[25] Z. Xu, S. Elshamy, and T. Fingscheidt, "A priori SNR estimation using discriminative non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 661–665.

[26] ITU, *Rec. P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*. International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Feb. 2001.

[27] ITU, *Rec. P.862.1: Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO*. International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Nov. 2003.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 4214–4217.

[29] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008.

[30] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[32] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, Feb. 1967.

[33] J. Abel, M. Strake, and T. Fingscheidt, "Artificial bandwidth extension using deep neural networks for spectral envelope estimation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Xi'an, China, Sep. 2016, pp. 1–5.

[34] J. Abel and T. Fingscheidt, "A DNN regression approach to speech enhancement by artificial bandwidth extension," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Dec. 2017, pp. 219–223.

[35] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 71–83, Jan. 2018.

[36] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Upper Saddle River, NJ, USA: Prentice Hall, Inc., 1987.

[37] F. Chollet *et al.*, Keras. (2015). Available: https://keras.io

[38] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software. [Online]. Available: https://www.tensorflow.org/

[39] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford University Press, Inc., 1995.

[40] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, vol. 9, Sardinia, Italy, May 2010, pp. 249–256.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] J. S. Garofolo *et al.*, *TIMIT aAcoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium (LDC), 1993.

[43] *Super Wideband Stereo Speech Database*. NTT Advanced Technology Corporation (NTT-AT).

[44] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010, pp. 3110–3113.

[45] ETSI, *EG 202 396-1: Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database*. European Telecommunications Standards Institute, Sep. 2008.

https://doi.org/10.24355/dbbs.084-202102121510-0

[46] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*. International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.
[47] S. Gustafsson, R. Martin, and P. Vary, "On the optimization of speech enhancement systems using instrumental measures," in *Proc. Workshop Quality Assessment Speech Audio Image Commun.*, Darmstadt, Germany, Mar. 1996, pp. 36–40.
[48] ITU, *Rec. P.1100: Narrow-Band Hands-Free Communication in Motor Vehicles*. International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Jan. 2015.
[49] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

Samy Elshamy received the B.Sc. degree in bioinformatics from the Friedrich-Schiller-Universität Jena, Germany, in 2011, and the M.Sc. degree in computer science from the Technische Universität Braunschweig, Germany, in 2013. He is currently working toward the Ph.D. degree in the field of speech enhancement at the Institute for Communications Technology, Technische Universität Braunschweig, Germany.

Tim Fingscheidt (S'93–M'98–SM'04) received the Dipl.-Ing. degree in electrical engineering in 1993 and the Ph.D. degree in 1998 from the RWTH Aachen University, Aachen, Germany.

He further pursued his work on joint speech and channel coding as a Consultant in the Speech Processing Software and Technology Research Department at AT&T Labs, Florham Park, NJ, USA. In 1999, he entered the Signal Processing Department of Siemens AG (COM Mobile Devices) in Munich, Germany, and contributed to speech codec standardization in ETSI, 3GPP, and ITU-T. In 2005, he joined Siemens Corporate Technology in Munich, Germany, leading the speech technology development activities in recognition, synthesis, and speaker verification. Since 2006, he is Full Professor with the Institute for Communications Technology, Technische Universität Braunschweig, Germany. His research interests include speech and audio signal processing, enhancement, transmission, recognition, and instrumental quality measures.

Dr. Fingscheidt received several awards; among them are a prize of the Vodafone Mobile Communications Foundation in 1999 and the 2002 prize of the Information Technology branch of the Association of German Electrical Engineers (VDE ITG). In 2017, he co-authored the ITG award-winning publication, which is awarded only once in a life time. He has been a speaker of the Speech Acoustics Committee ITG AT3 since 2015. From 2008 to 2010, he was an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and since 2011, he serves as a member of the IEEE Speech and Language Processing Technical Committee.

https://doi.org/10.24355/dbbs.084-202102121510-0

Publication VII

S. Elshamy and T. Fingscheidt, "Improvement of Speech Residuals for Speech Enhancement," in *Proc. of WASPAA*, New Paltz, NY, USA, Oct. 2019, pp. 214–218

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0

IMPROVEMENT OF SPEECH RESIDUALS FOR SPEECH ENHANCEMENT

Samy Elshamy and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, D–38106 Braunschweig, Germany

ABSTRACT

In this work we present two novel methods to improve speech residuals for speech enhancement. A deep neural network is used to enhance residual signals in the cepstral domain, thereby exceeding a former cepstral excitation manipulation (CEM) approach in different ways: One variant provides higher speech component quality by 0.1 PESQ points in low-SNR conditions, while another one delivers substantially higher noise attenuation by 1.5 dB, without loss of speech component quality or speech intelligibility. Compared to traditional speech enhancement based on the decision-directed (DD) *a priori* SNR estimation, a gain of even up to 3.5 dB noise attenuation is obtained. A semi-formal comparative category rating (CCR) subjective listening test confirms the superiority of the proposed approach over DD by 0.25 CMOS points (or even by 0.48 if two outlier subjects are not considered).

Index Terms— a priori SNR, speech enhancement, deep learning, cepstrum

1. INTRODUCTION

The aim of speech enhancement is to improve speech quality and speech intelligibility in speech communication. Traditional noise reduction algorithms consist of a noise power estimator [1, 2, 3], subsequent *a priori* SNR estimation [4, 5, 6, 7], and filtering with a spectral weighting rule [8, 9, 10], to obtain the enhanced speech signal. However, recent advances in deep learning have facilitated the use of more wholesome structures, among which several end-to-end solutions based on deep neural networks (DNN) have emerged [11, 12, 13]. However, as reported in [13], framewise enhanced signals often suffer from discontinuities. The framewise processing of speech signals is quite important for low-delay applications such as telephony, so that an enhancement on, e.g., utterance level is not feasible. A way to move from such approaches towards more generic and modular structures is proposed in [14]. Therein, the transition from traditional noise reduction to DNN-based approaches still relying on classical structures is nicely depicted. Recent advances in speech enhancement particularly include the learning of spectral weighting rules [15].

However, modularity can be further pushed by understanding the production of a speech signal after the source-filter model. For example, to combat reverberation, the residual signal obtained by linear predictive coding (LPC) analysis is weighted according to specific characteristics of reverberant speech signals and subsequently used for synthesis to reduce reverberation [16]. Furthermore, the classical spectral subtraction approach for noise reduction is augmented by modifying the residual signal obtained by LPC analysis [17]. The residual signal is weighted by a function which is calculated on the basis of the energy and also the kurtosis of the noisy speech signal. It is then used to excite the all-pole filter to synthesize the enhanced speech.

In this paper, we present a DNN-based version of the cepstral excitation manipulation approach (CEM) [7], which is used to estimate an improved *a priori* SNR for speech enhancement. We replace the classical signal processing-based manipulations of [7] by a regression approach based on a feedforward DNN, still operating in the cepstral domain (see also [18]). We also propose a variant that allows to obtain higher speech component quality, especially in the low-SNR conditions at the cost of less noise attenuation, still outperforming [7] in all metrics. We instrumentally evaluate all methods in a common speech enhancement framework, where we also report results for the traditional decision-directed (DD) *a priori* SNR estimation approach [8]. Additionally, a semi-formal comparison category rating (CCR) subjective listening test is conducted.

This paper is structured as follows: First, the system model and the baselines are shown in Section 2. Second, we introduce the new DNN-based CEM approaches in Section 3, followed by the experimental evaluation and the results of the listening test in Section 4. We finally conclude the paper in Section 5.

2. FRAMEWORK AND BASELINES

In the following, we will briefly introduce our signal model, the speech enhancement framework, and the baselines. We use an additive model where the time-domain microphone signal is defined as $y(n) = s(n) + d(n)$, with $s(n)$ being the clean speech signal and $d(n)$ the noise signal. The discrete-time sample index is denoted by n. The corresponding frequency-domain representation is obtained by applying a K-point discrete Fourier transform (DFT) as $Y_\ell(k) = S_\ell(k) + D_\ell(k)$. The frequency bin index is denoted by $0 \leq k \leq K-1$, and the frame index by ℓ. The signals are assumed to be statistically independent and having zero mean.

The speech enhancement framework for evaluation consists of the minimum statistics (MS) [1] noise power estimation algorithm to obtain $\hat{\sigma}_\ell^D(k)^2$, the *a priori* SNR estimator under test ($\hat{\xi}_\ell(k)$), and the minimum mean square error log-spectral amplitude (MMSE-LSA) estimator [8] to calculate the spectral weighting rule $G_\ell(k) = f(\hat{\xi}_\ell(k), \hat{\gamma}_\ell(k))$. Hereby, $\hat{\gamma}_\ell(k) = \frac{|Y_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}$ is the *a posteriori* SNR estimate. The weighting rule is finally applied to the microphone signal to obtain the enhanced speech as $\hat{S}_\ell(k) = G_\ell(k) \cdot Y_\ell(k)$. All time-domain signals are obtained after inverse DFT (IDFT) and subsequent synthesis by overlap-add.

2.1. Decision-Directed (DD) Baseline

As classical state-of-the-art approach we use the DD *a priori* SNR estimator [4]. The parameters are set to $\beta_{\mathrm{DD}} = 0.975$ and $\xi_{\min} = -15\,\mathrm{dB}$. Following, the *a priori* SNR is estimated as

$$\hat{\xi}_\ell^{\mathrm{DD}}(k) = (1-\beta_{\mathrm{DD}}) \cdot \max\{\hat{\gamma}_\ell(k)-1, 0\} + \beta_{\mathrm{DD}} \frac{|\hat{S}_{\ell-1}(k)|^2}{\hat{\sigma}_{\ell-1}^D(k)^2}.$$

https://doi.org/10.24355/dbbs.084-202102121510-0

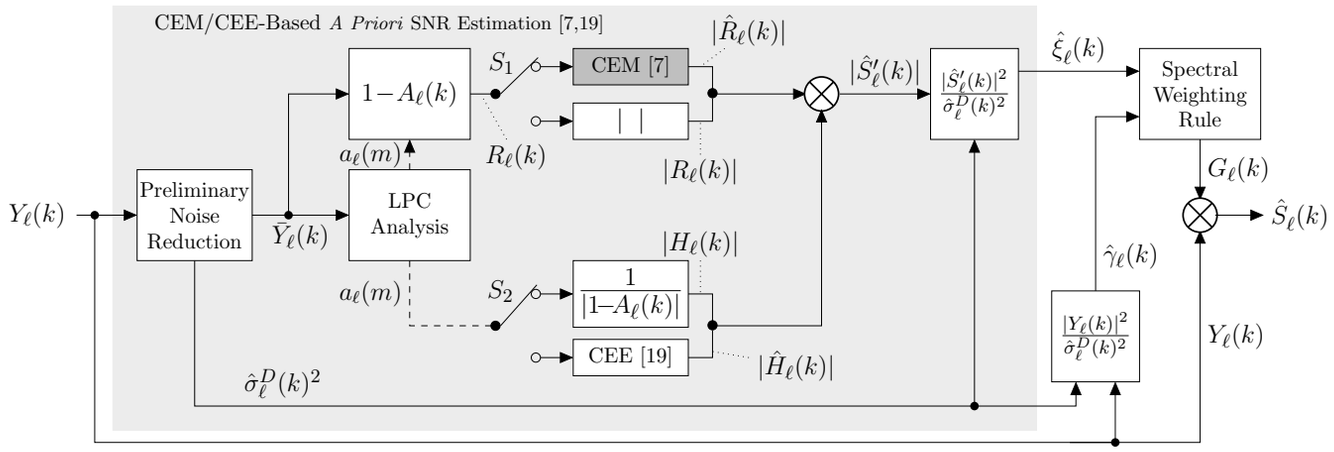


Figure 1: Schematic of the **two baseline** approaches **cepstral excitation manipulation (CEM)** [7] (switches S_1 and S_2 in upper position) and **cepstral envelope estimation (CEE)** [19] (switches S_1 and S_2 in lower position). The CEM block will be replaced by our **newly proposed CEM-DNN**, as detailed in Figure 2.

2.2. Cepstral Excitation Manipulation (CEM) Baseline

The schematic of the CEM approach is depicted in Figure 1, when switches S_1 and S_2 both are in upper position. The preliminary noise reduction consists of the MS noise power estimator, the DD *a priori* SNR estimator, and the MMSE-LSA spectral weighting rule. Please note that it is not restricted to that specific setup but we have made good experiences with it. It processes $Y_\ell(k)$ to obtain a more suitable signal $\bar{Y}_\ell(k)$ for the subsequent processing pipeline. Next, the enhanced signal $\bar{Y}_\ell(k)$ is subject to LPC analysis. The signal is then split into its excitation in the upper path, and its envelope in the lower path which is left untouched in this case. The excitation signal is manipulated according to [7] as follows. First, the signal is transformed from the frequency domain to the cepstral domain by applying a discrete cosine transform of type II (DCT-II) to the logarithmized amplitudes of $R_\ell(k)$, to obtain the coefficients $c_\ell^R(m)$ with m being the quefrency bin index. Second, a pitch estimation algorithm [20] is used to identify the quefrency bin index m_{F_0} with the highest amplitude in a certain range of bins, that represent typical fundamental frequencies. Third, according to m_{F_0}, a pretrained excitation template from clean speech $c_\ell^{\hat{R}}(m)$ is used for further processing. The template's energy is then adjusted to match the energy of the preliminary denoised residual $c_\ell^R(m)$ as

$$c_\ell^{\hat{R}}(0) = c_\ell^R(0). \tag{1}$$

Furthermore, the amplitude of the quefrency bin that is corresponding to the fundamental frequency is overestimated to boost the harmonic structure and cause an attenuation of the noise between the harmonics by

$$c_\ell^{\hat{R}}(m_{F_0}) = \alpha \cdot c_\ell^R(m_{F_0}). \tag{2}$$

The overestimation factor is chosen as $\alpha > 1$. The manipulated vector $c_\ell^{\hat{R}}(m)$ is then transformed back by an inverse DCT-II and finally some start and end decay is applied to simulate a more steady rise and decay of the excitation signal. Finally, the enhanced residual signal $|\hat{R}_\ell(k)|$ is mixed with the preliminary denoised spectral envelope $|H_\ell(k)|$, and used further as new clean speech amplitude estimate for the *a priori* SNR estimate. For further details we kindly refer to [7].

2.3. Cepstral Envelope Estimation (CEE) Baseline

The counterpart of the CEM approach is the CEE method which is also depicted in Figure 1, when switches S_1 and S_2 are both in lower postition. Here, the aim is to enhance the spectral envelope separately from the excitation signal by estimating a clean spectral envelope $|\hat{H}_\ell(k)|$ corresponding to the preliminary denoised observation. This is done after LPC analysis in the lower path of Figure 1. A classification feedforward DNN with six hidden layers, each with 58 nodes and sigmoid activation functions is used for the enhancement [19]. The $N = 10$ LPC coefficients are converted to cepstral coefficients by [21] in order to circumvent filter instabilities. They are then used as input features for the DNN and are subsequently mapped to a probability distribution over 65 potential spectral envelope prototypes, which are stored in a codebook as cepstral coefficients. This allows to calculate an MMSE estimate over all prototypes, according to their respective probability as delivered by the DNN. Finally, the estimate is transformed back to the spectral domain to obtain $|\hat{H}_\ell(k)|$, and used along with the preliminary denoised residual signal $|R_\ell(k)|$ in the numerator of the *a priori* SNR estimate. One drawback of the spectral envelope enhancement is that it is unable to further improve the noise attenuation between the harmonics, which is then limited by $|R_\ell(k)|$. However, the CEE method will be used as a baseline and also in serial concatenation (denoted by the $\rightarrow$ symbol) with the CEM approaches to alleviate this drawback. Further details, especially on the concatenation, can be found in [19].

3. NEW DNN-BASED CEPSTRAL EXCITATION MANIPULATION (CEM-DNN)

As mentioned in Section 1, we want to use DNNs in a more modular fashion to enhance the speech residual, as this has already shown good success for the enhancement of the spectral envelope [19]. Thus, we also want to replace the earlier used templates as mentioned in Section 2.2, and the manipulations from (1) and (2) by a regression DNN as shown in Figure 2. The first step is a feature conversion, where the spectral representation of the residual signal $R_\ell(k)$ is transformed to the cepstral domain by application of a DCT-II to obtain $c_\ell^R(m)$. The cepstral excitation is then subject

https://doi.org/10.24355/dbbs.084-202102121510-0

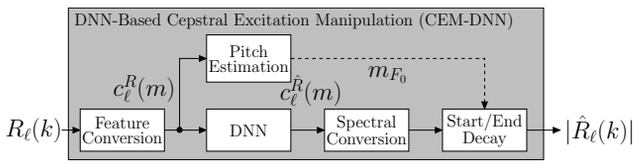


Figure 2: Diagram of the **novel DNN-based cepstral excitation manipulation (CEM-DNN)** approach that uses a DNN instead of the classical manipulations.

to the simple pitch estimator from [20], which is no longer required for the manipulations as in Section 2.2, but only for the start and end decay block which ensures a steady rise and decay of the generated excitation signal. Parallel to that, the actual manipulations are done by the DNN, which is followed by a spectral conversion where an IDCT-II is used to transform the signal back to the spectral domain. Finally, some start and end decay is applied, as before. This approach is referred to as **CEM-DNN**.

The feedforward DNN comprises six hidden layers, 512 nodes each and sigmoid activation functions. The input and output layers have 256 nodes each and the network is trained to perform regression. It has shown to be important to normalize not only the input features but also the target representation, in order to preserve harmonic structures. Hence, all data is mean- and variance-normalized according to the respective dataset[1].

In this work, we also investigate the replacement of the formerly used templates and the overestimation (2) by a DNN, but still applying (1) on the network's rescaled output. Thereby, we hope to also improve the speech component quality and speech intelligibility. To allow for a better speech component quality, the strong noise attenuation might be alleviated to some extent by retaining the energy coefficient. This is denoted as **CEM-DNN-$\mathbf{c_0}$**.

4. EXPERIMENTAL EVALUATION

4.1. Setup

All experiments are conducted at a sampling rate of 8 kHz. We use a periodic square root Hann window for analysis and synthesis with a frame length of $K = 256$ and a frame shift with 50 % overlap. Utilized clean speech databases are the TIMIT database [22] for training and development (relevant for Section 3) and the NTT super-wideband database [23], where only English speakers are used, for testing. Two noise databases are utilized, QUT [24] and ETSI [25], where the first is only for training and development, and the second only for testing. Four noise files[2] are used exclusively for an unseen test set. We simulate six SNR conditions from -5 dB to 20 dB in steps of 5 dB. Levels are measured according to ITU-T P.56 [26], adjusted, and subsequently superimposed to obtain the noisy observations.

4.2. Metrics

For instrumental evaluation of the approaches we investigate three different metrics. Two metrics operate on either the so-called *filtered* clean speech component $\tilde{s}(n)$ or the *filtered* noise component $\tilde{d}(n)$. Both are obtained by employing the white-box approach [27], meaning that the gain function $G_\ell(k)$ of the speech enhancement system is not only applied to $Y_\ell(k)$, but also to the separate components $S_\ell(k)$ and $D_\ell(k)$. After IDFT and subsequent overlap-add synthesis the *filtered* components are obtained. We measure the segmental noise attenuation (NA_{seg}) [28] as

$$\text{NA}_{\text{seg}} = 10\log_{10}\left[\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\text{NA}(\ell)\right], \tag{3}$$

with

$$\text{NA}(\ell) = \frac{\sum_{\nu=0}^{N-1} d(\nu+\ell N)^2}{\sum_{\nu=0}^{N-1}\tilde{d}(\nu+\ell N)^2}.$$

Please note that both signals have to be time-aligned. Here, $\ell \in \mathcal{L}$ represents all frames, where each frame contains $N = 256$ samples.

Furthermore, we assess the so-called speech *component* quality by measuring the PESQ score [29] of the *filtered* speech component $\tilde{s}(n)$, dubbed PESQ($\tilde{s}$), where the clean speech component $s(n)$ serves as reference. We measure PESQ on the *filtered* speech component since it has not been validated for artifacts that might be introduced by noise reduction algorithms. The speech intelligibility is measured on the enhanced signal $\hat{s}(n)$ by the short-time objective intelligibility measure (STOI) [30].

4.3. Instrumental Evaluation

We show the results[3] in Figs. 3 and 4, where each marker represents one SNR condition, starting at the bottom with -5 dB, up to 20 dB at the top. The results show that the **CEM-DNN** approach (square markers, solid green line) already outperforms all baselines in terms of NA_{seg}. Only in some high-SNR conditions, where PESQ($\tilde{s}$) is already very high (more than 4 PESQ points), the speech component quality of the baselines outperforms **CEM-DNN**. Even though the **CEE** approach (circle markers, solid orange line) shows convincing performance, it still lacks attenuation between the harmonics, as earlier mentioned. This is also reflected in Figure 4, where no significant improvement in speech intelligibility is seen. The noise attenuation can be further increased by the serial concatenation of CEE and CEM-DNN resulting in **CEE** $\rightarrow$ **CEM-DNN** (triangle markers, solid green line), which obtains up to 1.5 dB higher NA_{seg} compared to **CEE** $\rightarrow$ **CEM** (triangle markers, dashed green line) and even 3.5 dB over **DD**, both, without significant loss of speech component quality or speech intelligibility.

The alternative approaches **CEM-DNN-$\mathbf{c_0}$** (square markers, solid purple line) and **CEE** $\rightarrow$ **CEM-DNN-$\mathbf{c_0}$** (triangle markers, solid purple line), which retain the cepstral energy coefficient from the preliminary denoised residual signal, show less NA_{seg} compared to their corresponding analogues which also estimate the energy coefficient. However, they still outperform their corresponding baselines w.r.t. noise attenuation, and absolute improvement over **CEE** $\rightarrow$ **CEM** is obtained. In addition to that, STOI is slightly improved in the important -5 dB SNR condition, and PESQ($\tilde{s}$) is improved by more than 0.1 points compared to the other **CEM**-based approaches and also **DD**.

4.4. Subjective Listening Test

Finally, we conduct a semi-formal CCR listening test [31, Annex E]. The subjects are always presented two conditions A and B,

[1] This is also feasible in a practical system, as the algorithm operates in the uplink and allows to maintain statistics for the network's input and also for the rescaling of the network's output. It is sufficient to obtain the required statistics for rescaling of the network's output once in a while in assumed good SNR conditions.

[2] `Fullsize Car1 80Kmh, Outside Traffic Crossroads, Pub Noise Binaural V2, Work Noise Office Callcenter`

[3] Audio samples can be found under: https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2019-waspaa-cem/

https://doi.org/10.24355/dbbs.084-202102121510-0

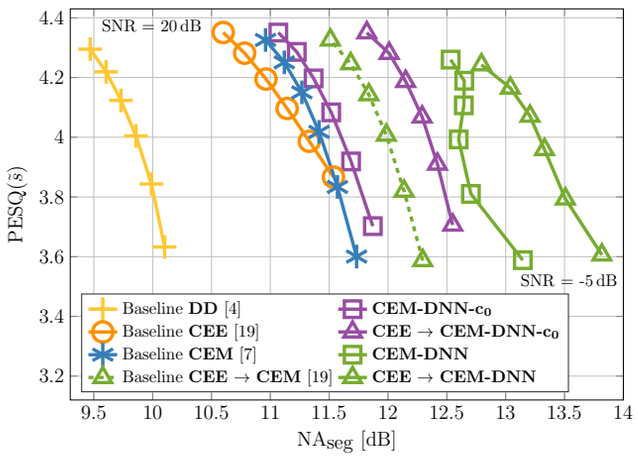


Figure 3: Speech component quality (**PESQ**($\tilde{s}$)) vs. noise attenuation (NA$_{\text{seg}}$) in unseen and also non-stationary noises.

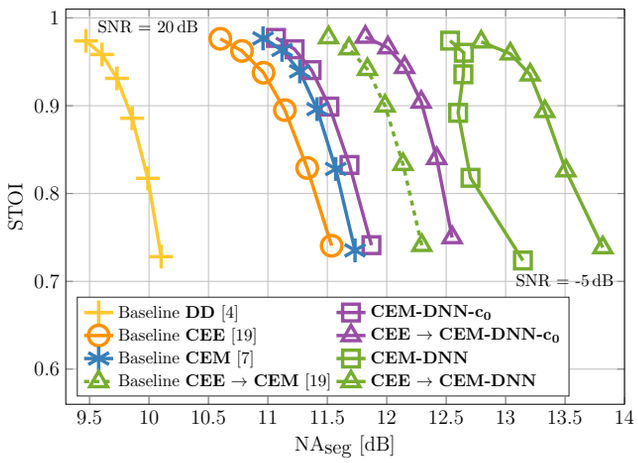


Figure 4: Speech intelligibility (**STOI**) vs. noise attenuation (NA$_{\text{seg}}$) in unseen and also non-stationary noises

Table 1: CMOS results and 95 % confidence intervals for the subjective listening test. The preferred approach is in **boldface**.

Condition	CMOS	CI_{95}
a) Noisy vs. **DD**	0.96	± 0.16
b) Noisy vs. **CEM-DNN**	0.97	± 0.22
c) Noisy vs. **CEE → CEM-DNN**	1.22	± 0.20
d) DD vs. **CEE → CEM-DNN**	**0.25**	± 0.17
e) CEM-DNN vs. **CEE → CEM-DNN**	0.12	± 0.11
f) CEE → CEM-DNN-c$_0$ vs. **CEE → CEM-DNN**	0.07	± 0.13

and are subsequently asked to judge the relative quality of B over A based on the comparative mean opinion score (CMOS) scale. The CMOS scale ranges from -3 (much worse) in integer steps to 3 (much better). The presented samples are based on a female and a male German speaker's sentence from the NTT database [23], more test conditions would have extended the subjective test too much. They are superimposed with two noises[4] from the ETSI database [25] at -5 dB and 5 dB SNR. All files are upsampled to 48 kHz for replay. In total six conditions are presented: a) Noisy vs. **DD**, b) Noisy vs. **CEM-DNN**, c) Noisy vs. **CEE → CEM-DNN**, d) **DD** vs. **CEE → CEM-DNN**, e) **CEM-DNN** vs. **CEE → CEM-DNN**, and f) **CEE → CEM-DNN-c$_0$** vs. **CEE → CEM-DNN**. This amounts to 2 (SNRs) × 2 (sentences) × 2 (noises) × 6 (conditions) = 48 comparisons. As they have to be presented in both directions, A followed by B and vice versa, the total amount is 96. The samples are played back to the subjects via `AKG K-271 MKII` headphones, attached to `Fireface` devices, connected through `FireWire` to a standard personal computer.

The results for the CCR listening test including all subjects are depicted in Table 1, where CMOS and 95 % confidence intervals (CI_{95}) are shown. In total 17 native speakers participated in the subjective listening test, where two outliers strongly preferred the noisy reference for conditions a), b), and c). Despite the two outliers the first three rows show that each of the tested approaches (**DD**, **CEM-DNN**, and **CEE → CEM-DNN**) gain about 1 CMOS point if compared to the noisy condition, with **CEE → CEM-DNN** performing best by obtaining 1.22 CMOS points. The various CEM-based approaches compared among themselves do not show a significant preference of the listeners. However, the **CEM-DNN** approach obtains significant 0.25 CMOS points when compared to the **DD** approach, which indicates that the high gain in noise attenuation is positively rewarded by the listeners.

[4] `Outside Traffic Crossroads, Work Noise Office Callcenter`

Table 2: CMOS results and 95 % confidence intervals for the subjective listening test with two outlier subjects not considered. The preferred approach is in **boldface**.

Condition	CMOS	CI_{95}
a) Noisy vs. **DD**	1.19	± 0.16
b) Noisy vs. **CEM-DNN**	1.35	± 0.19
c) Noisy vs. **CEE → CEM-DNN**	1.62	± 0.16
d) DD vs. **CEE → CEM-DNN**	**0.48**	± 0.16
e) CEM-DNN vs. **CEE → CEM-DNN**	0.15	± 0.11
f) CEE → CEM-DNN-c$_0$ vs. **CEE → CEM-DNN**	0.17	± 0.14

In Table 2, where the two outlier subjects who preferred the noisy conditions have not been considered, the results are even more significant. Even condition f) obtains a significant result that indicates a preference of the subjects for higher noise attenuation over higher speech component quality. Especially the preference of the new **CEM-DNN** over **DD** by 0.48 CMOS points is quite distinct.

5. CONCLUSIONS

In this work we have presented a novel method that incorporates deep neural networks (DNNs) to enhance speech residuals in the cepstral domain for speech enhancement. Two variants of the DNN-based cepstral excitation manipulation (CEM) approach are investigated: One focuses on the speech component quality and gains more than 0.1 PESQ points in the -5 dB SNR condition compared to the former classical CEM and also the decision-directed (DD) approach. The other obtains substantially higher noise attenuation by up to 1.5 dB over CEM and even 3.5 dB over DD. A semi-formal comparative category rating (CCR) listening test has shown that the serial combination of cepstral envelope enhancement and DNN-based CEM is better by 0.25 CMOS points (or even 0.48 when two outlier subjects are not considered) than the classical DD approach.

https://doi.org/10.24355/dbbs.084-202102121510-0

6. REFERENCES

[1] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE T-SAP*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[2] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE T-SAP*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[3] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay," *IEEE T-ASLP*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE T-ASSP*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[5] C. Breithaupt, T. Gerkmann, and R. Martin, "A Novel A Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing," in *Proc. of ICASSP*, Las Vegas, NV, USA, Mar. 2008, pp. 4897–4900.

[6] S. Suhadi, C. Last, and T. Fingscheidt, "A Data-Driven Approach to A Priori SNR Estimation," *IEEE T-ASLP*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[7] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous A Priori SNR Estimation by Cepstral Excitation Manipulation," *IEEE/ACM T-ASLP*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[8] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE T-ASSP*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[9] P. Scalart and J. V. Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP*, Atlanta, GA, USA, May 1996, pp. 629–632.

[10] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM T-ASLP*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[12] Y. Wang and D. L. Wang, "A Deep Neural Network for Time-Domain Signal Reconstruction," in *Proc. of ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4390–4394.

[13] S.-W. Fu, T.-W. Wang, Y. Taso, X. Lu, and H. Kawai, "End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks," *IEEE/ACM T-ASLP*, vol. 26, no. 9, pp. 1570–1584, Sept. 2018.

[14] S. Mirsamadi and I. Tashev, "Causal Speech Enhancement Combining Data-Driven Learning and Suppression Rule Estimation," in *Proc. of Interspeech*, San Francisco, CA, USA, Sept. 2016, pp. 2870–2874.

[15] Y. Wang, A. Narayanan, and D. L. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM T-ASLP*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[16] B. Yegnanarayana and P. S. Murthy, "Enhancement of Reverberant Speech Using LP Residual Signal," *IEEE T-ASSP*, vol. 8, no. 3, pp. 267–281, May 2000.

[17] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of Noisy Speech by Spectral Subtraction and Residual Modification," in *Proc. of INDICON*, New Delhi, India, Sept. 2006.

[18] S. Elshamy and T. Fingscheidt, "DNN-Based Cepstral Excitation Manipulation for Speech Enhancement," Submitted to *IEEE/ACM T-ASLP*.

[19] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "DNN-Supported Speech Enhancement With Cepstral Estimation of Both Excitation and Envelope," *IEEE/ACM T-ASLP*, vol. 26, no. 12, pp. 2460–2474, Dec. 2018.

[20] A. M. Noll, "Cepstrum Pitch Determination," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, Feb. 1967.

[21] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Upper Saddle River, NJ, USA: Prentice Hall, Inc., 1987.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium (LDC), 1993.

[23] "Super Wideband Stereo Speech Database," NTT Advanced Technology Corporation (NTT-AT).

[24] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT Corpus for the Evaluation of Voice Activity Detection Algorithms," in *Proc. of Interspeech*, Makuhari, Japan, Sept. 2010, pp. 3110–3113.

[25] ETSI, *EG 202 396-1: Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database*, European Telecommunications Standards Institute, Sept. 2008.

[26] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.

[27] S. Gustafsson, R. Martin, and P. Vary, "On the Optimization of Speech Enhancement Systems Using Instrumental Measures," in *Proc. of Workshop on Quality Assessment in Speech, Audio, and Image Communication*, Darmstadt, Germany, Mar. 1996, pp. 36–40.

[28] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-Optimized Speech Enhancement," *IEEE T-ASLP*, vol. 16, no. 4, pp. 825–834, May 2008.

[29] ITU, *Rec. P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Feb. 2001.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE T-ASLP*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[31] ITU, *Rec. P.800: Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Aug. 1996.

https://doi.org/10.24355/dbbs.084-202102121510-0

https://doi.org/10.24355/dbbs.084-202102121510-0